UNIVERSIDAD AUTONOMA DE MADRID

Sistematización de la metodología de validación e interpretación de resultados para medidas clínicas de la enfermedad de Parkinson

Tesis Doctoral realizada por

Mª del Carmen Rodríguez Blázquez

Bajo la dirección de

Dr. Pablo Martínez Martín

Departamento de Medicina Preventiva, Salud Pública y Microbiología
Facultad de Medicina
Universidad Autónoma de Madrid

El Dr. Pablo Martínez Martín, director de la tesis doctoral y Científico Titular del Centro Nacional de Epidemiología (Instituto de Salud Carlos III)

INFORMA:

Que Dª. María del Carmen Rodríguez Blázquez ha realizado bajo su dirección el estudio titulado *Sistematización de la metodología de validación e interpretación de resultados para medidas clínicas de la enfermedad de Parkinson*. Es un trabajo original, rigurosamente realizado y es apto para ser defendido públicamente con el fin de obtener el grado de Doctor.

Para que así conste y surta los efectos oportunos, se firma este documento en Madrid, a 12 de diciembre de 2014.

Fdo. Dr Pablo Martínez Martín

A mis padres.

A mis hijos.

Agradecimientos

A Pablo Martínez Martín, por su insuperable labor como Director de esta Tesis, y por haberme ayudado y aportado tanto personal y profesionalmente con generosidad y paciencia. Muchas gracias, Milord.

A Jesús de Pedro Cuesta, que me dio la oportunidad de trabajar en su grupo, por su profesionalidad, dedicación y calidad humana.

A Maria João Forjaz, compañera y amiga, con quien comparto tantos proyectos, momentos y confidencias a lo largo del día, por su constante apoyo y por haberme enseñado tanto.

A Fernando Rodríguez Artalejo y al Departamento de Medicina Preventiva y Salud Pública y Microbiología de la Universidad Autónoma de Madrid, por las facilidades dadas y la amabilidad con que me han tratado.

A mis compañeros del Centro Nacional de Epidemiología, en especial a Fuencisla, Javier, Quique, Mar, Carmen, Rocío, Paloma, Oliva, y a Fermi, Gloria, Belén, Alba y Maru, por los cafés, por los buenos ratos y por estar siempre cerca.

A los neurólogos y pacientes que han aportado los datos que han hecho posible las publicaciones y estudios de esta Tesis.

A mi familia, por existir y hacer del mundo un lugar mejor. Y sobre todo, a mi padre, por ser la luz que me acompañará siempre.

Resumen

Antecedentes

Ante una entidad tan compleja como la enfermedad de Parkinson, es necesario contar con instrumentos de evaluación que ayuden a identificar la presencia de síntomas y complicaciones y su gravedad e impacto, documenten la evolución del proceso, valoren el efecto de las intervenciones terapéuticas y faciliten el intercambio de información entre clínicos, con otros profesionales y con los propios pacientes y sus familias. En las últimas décadas se han desarrollado una gran variedad de escalas para la evaluación de las manifestaciones clínicas de la enfermedad de Parkinson. Paralelamente, se han publicado importantes directrices para la validación de escalas clínicas y de resultados informados por los pacientes, que definen los requisitos que debe cumplir un instrumento de medición del estado de salud para garantizar su calidad. En la actualidad, existe un consenso generalizado acerca de que la creación y desarrollo de una medida del estado de salud debe tener en cuenta una serie de principios y seguir una metodología muy precisa que proviene del campo de la Psicología y la Psicometría.

En este contexto, la determinación de los principales atributos de una medida de la salud (aceptabilidad, fiabilidad, validez y precisión) y su comparación con valores aceptados como umbrales de calidad han supuesto una aportación central de los trabajos recogidos en esta Tesis. A esta "sistemática de análisis" se han incorporado métodos de determinación de sensibilidad al cambio y su interpretación, de gran trascendencia clínica, aunque carente de estándares ampliamente aceptados.

Objetivos

Analizar las propiedades métricas de varios instrumentos de medida de la salud (escalas clínicas y resultados comunicados por los pacientes) para enfermedad de Parkinson, y analizar e interpretar el cambio en una cohorte de pacientes con enfermedad de Parkinson utilizando algunas de dichas medidas.

Material y métodos

La Tesis presenta los resultados del análisis de las propiedades métricas de las siguientes escalas clínicas y de resultados comunicados por los pacientes con enfermedad de Parkinson: *Clinical Impression of Severity Index – Parkinson's Disease* (CISI-PD), *Hospital Anxiety and Depression Scale* (HADS), *Non-Motor Symptoms Scale* (NMSS), *Scales for Outcomes in Parkinson's Disease – Autonomic* (SCOPA-AUT), *modified Parkinson Psychosis Rating Scale* (mPPRS), y *Movement Disorders Society sponsored version of the Unified Parkinson's Disease Rating Scale* (MDS-UPDRS). Se ha aplicado una metodología de validación sistematizada a partir de una amplia revisión bibliográfica de las directrices existentes sobre el tema. Las propiedades métricas analizadas son: viabilidad y aceptabilidad, fiabilidad, validez y preción. Para cada una de ellas, se ha comprobado su adecuación a los criterios de calidad estándares, presentándose los resultados de los artículos que componen esta Tesis. Con los resultados de seguimiento del Estudio Longitudinal de Enfermedad de Parkinson (ELEP) se ha analizado la sensibilidad al cambio e interpretabilidad de las puntuaciones de las escalas CISI-PD, HADS, SCOPA-AUT y mPPRS mediante técnicas basadas en la distribución de las puntuaciones.

Resultados:

Todas las escalas, excepto la mPPRS, mostraron una viabilidad y aceptabilidad satisfactorias. En cuanto a la fiabilidad entendida como consistencia interna, se observó que las puntuaciones totales de las escalas analizadas mostraron una buena consistencia interna, aunque no fue así en el caso de algunas subescalas de la NMSS, SCOPA-AUT y MDS-UPDRS. La fiabilidad test-retest se estudió en las escalas CISI-PD, NMSS y MDS-UPDRS, con resultados que apoyan la estabilidad de las puntuaciones. La validez de contenido solo se estudió formalmente en el caso de la mPPRS. La mayor parte de las escalas mostraron una validez de constructo o comprobación de hipótesis satifactoria. Los resultados del error estándar de medida fueron inferior al criterio, excepto en la mPPRS, por lo que se puede concluir que la precisión fue satisfactoria. En general, la evaluación global de las escalas fue coincidente con el grado de recomendación que ofrecen las respectivas revisiones de la *Movement Disorders Society*.

Las puntuaciones de las escalas CISI-PD, HADS, mPPRS y SCOPA-AUT mostraron una tendencia al empeoramiento en el seguimiento a tres años realizado a los pacientes, si bien las diferencias sólo fueron estadísticamente significativas en las escalas CISI-PD y SCOPA-AUT. Los estadísticos tamaño del efecto y del cambio relativo arrojaron valores bajos en general. El umbral de mínimo cambio importante que permite clasificar a los pacientes que mejoran, permanecen estables y empeoran se calculó a partir del error estándar de la medida, ½ desviación típica basal y el 10% de la puntuación total de la escala y el valor estimado del cambio resultante de la triangulación (promedio) de las tres medidas. Como resultado, se detectó un empeoramiento en el 14%-38% de los pacientes, especialmente en las escalas CISI-PD y SCOPA-AUT.

Conclusiones:

De acuerdo con la metodología de validación de escalas de medida clínica y de resultados comunicados por los pacientes presentada en esta Tesis, las escalas analizadas resultaron ser aceptables y viables, fiables, válidas y precisas, por lo que resultan de utilidad para la práctica clínica y la investigación. A pesar de que los valores de sensibilidad al cambio resultaron bajos en general, se puede concluir que las escalas analizadas pueden detectar cambios debido a la evolución de la enfermedad y dichos cambios pueden dotarse de un significado clínico.

Palabras clave:

Evaluación, validación de escalas, enfermedad de Parkinson, sensibilidad al cambio, interpretabilidad.

Abreviaturas

CCI	Coeficiente de correlación intraclase
CISI-PD	*Clinical Impression of Severity Index – Parkinson's Disease*
COSMIN	*Consensus-based Standards for the selection of health Measurement Instruments*
CR	Cambio relativo
DT	Desviación típica
EEM	Error estándar de la medida
ELEP	Estudio Longitudinal en Enfermedad de Parkinson
EP	Enfermedad de Parkinson
HADS	*Hospital Anxiety and Depression Scale*
HY	Clasificación de Hoehn y Yahr
MCI	Mínimo cambio importante
MDI	Mínima diferencia importante
MDS-UPDRS	*Movement Disorders Society sponsored version of the Unified Parkinson's Disease Rating Scale*
mPPRS	*Modified Parkinson Psychosis Rating Scale*
NMSS	*Non-Motor Symptoms Scale*
PRO	*Patient-reported outcomes*
RME	Respuesta media estandarizada
SCOPA	*Scales for Outcomes in Parkinson's Disease*
TAC	Tasa anual de cambio
TE	Tamaño del efecto
VEC	Valor estimado del cambio

Índice

Sistematización de la metodología de validación e interpretación de resultados para medidas clínicas de la enfermedad de Parkinson

INTRODUCCIÓN

1. Introducción

1.1. La enfermedad de Parkinson

La enfermedad de Parkinson (EP), descrita por primera vez por James Parkinson en 1817, es un trastorno neurológico progresivo que tradicionalmente se ha caracterizado por manifestaciones cardinales de índole motor: lentitud en los movimientos o bradicinesia, temblor, rigidez e inestabilidad postural. En general, la EP se diagnostica a partir de la presencia de bradicinesia acompañada de alguno de los otros tres signos cardinales, generalmente asimétricos en expresión, la respuesta al tratamiento con levodopa y la ausencia de signos "atípicos" que sugieran una enfermedad distinta a la EP (parálisis supranuclear progresiva, atrofia multisistémica, etc.) [1]. La presencia de trastornos de la marcha y del equilibrio apoyan el diagnóstico clínico, aunque algunas modalidades de estas manifestaciones (por ejemplo, las congelaciones y la pérdida de los reflejos posturales) solo son apreciables cuando la enfermedad ya está evidentemente desarrollada. Se trata de una enfermedad cuyos síntomas suelen comenzar en la edad adulta, hacia los 50-60 años, y que progresa lentamente durante alrededor de 10-20 años. Es la segunda enfermedad neurodegenerativa más frecuente, por detrás de la enfermedad de Alzheimer, con una prevalencia estimada en España, en población de 65 años o superior, entre el 1,1 y el 1,5% [2–5].

Tradicionalmente se ha considerado que la causa de la EP es la degeneración de las neuronas dopaminérgicas nigoestriatales, lo que explicaría los síntomas motores. Sin embargo, la afectación neurodegenerativa de otros circuítos (por ejemplo, mesolímbico), otras estructuras neurales (por ejemplo, el nucleus coeruleus) y, consecuentemente, otros neurotransmisores distintos a la dopamina (por ejemplo, acetilcolina) está ampliamente reconocida. Estas alteraciones justifican la consideración de la EP como una enfermedad compleja que afecta múltiples componentes del sistema nervioso central y explican el amplio espectro de manifestaciones "no motoras" que pueden, incluso, preceder a los síntomas motores, hechos que han llevado a algunos

autores a proponer una redefinición de la EP [6]. En este sentido, existiría una fase pre-clínica de la EP, en la que los pacientes no manifiestan síntomas ni signos de enfermedad, pero en la que es posible detectar la presencia de marcadores moleculares y de imagen en el sistema nervioso central (por ejemplo, bajos niveles de α-sinucleína en el líquido cerebroespinal), que podrían predecir la aparición de la EP. En la fase pre-motora emergen manifestaciones no motoras de la enfermedad, tales como fatiga, cambios conductuales y del estado de ánimo, y disfunción autonómica. En esta fase es de especial interés la presencia de síntomas tales como las alteraciones del sentido del olfato (hiposmia), estreñimiento, reducción de la variabilidad de la frecuencia cardíaca y síntomas de ansiedad y depresión, que se ha demostrado pueden predecir la aparición de la EP en varios años [7,8]. Por último, en la fase motora aparecen los signos cardinales clásicos descritos por James Parkinson (temblor, rigidez y lentitud en los movimientos o bradicinesia), que se acompañan de una amplia variedad de síntomas no motores: trastornos neuropsiquiátricos, deterioro cognitivo, alteraciones del sueño, síntomas autonómicos, fatiga, dolor, etc. que están presentes en más del 90% de los pacientes [9,10] (Tabla 1.1). El conjunto de manifestaciones y complicaciones motoras y no motoras repercute intensamente sobre la capacidad funcional del paciente y su ajuste psicosocial, siendo responsables de un considerable deterioro de su calidad de vida [10,11], del aumento de las tasas de hospitalización e institucionalización, con un elevado gasto socio-sanitario, y de la mortalidad [12–14].

Tabla 1.1. Principales síntomas motores y no motores en la enfermedad de Parkinson (adaptado de Chaudhuri y colaboradores, 2006 y Jankovic, 2008).

Síntomas motores
- Síntomas cardinales: temblor, bradicinesia, rigidez, inestabilidad postural. - Hipomimia - Disartria - Festinación - Micrografía - Blefarospasmo - Distonia - Camptocormia
Síntomas no motores
Síntomas cognitivos y neuropsiquiátricos
- Depresión, apatía, ansiedad - Anhedonia - Déficit de atención - Alucinaciones, delirios - Demencia - Conducta obsesiva (normalmente inducida por la medicación)
Trastornos del sueño
- Síndrome de piernas inquietas y movimientos periódicos de los miembros - Trastorno de conducta en fase REM, pérdida de atonía en fase REM - Somnolencia diurna excesiva - Ensoñaciones vívidas - Insomnio
Síntomas autonómicos
- Alteraciones urinarias: urgencia, nocturia, aumento de la frecuencia urinaria - Sudoración excesiva - Hipotensión ortostática - Disfunción sexual, hipersexualidad, disfunción eréctil - Sequedad en los ojos (xerostomía)
Síntomas gastrointestinales
- Babeo - Disfagia y atragantamientos - Reflujo, vómitos - Nausea - Estreñimiento y problemas al defecar (vaciamiento incompleto, incontinencia fecal)
Síntomas sensoriales
- Dolor - Parestesia - Trastornos olfativos
Otros síntomas
- Fatiga - Diplopia - Visión borrosa - Seborrea - Pérdida o ganancia de peso no relacionada con la dieta

1.2. Evaluación de la enfermedad de Parkinson

El diagnóstico y tratamiento de las manifestaciones clínicas de la enfermedad de Parkinson requiere una evaluación cuidadosa de las mismas para conocer su alcance y sus efectos en la vida del paciente. Para ello, se emplea una variedad de instrumentos de evaluación que permiten identificar la presencia de déficits y complicaciones, documentar la evolución del proceso, cuantificar la gravedad e impacto de los síntomas, valorar el efecto de las intervenciones terapéuticas, facilitar el intercambio de información entre clínicos, con otros profesionales y con los agentes de la política sociosanitaria, y con los propios pacientes y sus familias. Hasta hace dos décadas, se utilizaban escalas "ad hoc" confeccionadas por los clínicos, de las que se carecía de información básica sobre sus propiedades métricas. En la actualidad, existe un consenso generalizado acerca de que la creación y desarrollo de una medida del estado de salud debe tener en cuenta una serie de principios y seguir una metodología muy precisa que provienen del campo de la Psicología (psicometría), Sociología (análisis de indicadores sociales), y Pedagogía (interpretación de resultados de exámenes y pruebas).

En la enfermedad de Parkinson, la evaluación de las manifestaciones y complicaciones motoras y no motoras se realiza mediante escalas multi-dominio como la *Unified Parkinson's Disease Rating Scale* (UPDRS) o su versión más reciente, la *Movement Disorders Society-sponsored version* de la *UPDRS* (MDS-UPDRS) [15,16]; escalas de valoración global de la enfermedad, como la clasificación en estadíos de Hoehn y Yahr [17] y la *Clinical Impression of Severity Index-Parkinson's Disease* (CISI-PD); instrumentos de evaluación de los síntomas motores y el estado funcional, como la *Scales for Outcomes in Parkison's Disease* (SCOPA)-Motor [18], y la escala de Schwab & England (SES) [19]; y escalas de evaluación de las complicaciones y fluctuaciones motoras, tales como la serie *Wearing-off Questionnaires* (WOQ) [20–22] y la *Unified Dyskinesia Rating Scale* (UDysRS) [23]. Se trata de escalas ampliamente usadas en diversos ámbitos (clínico, investigación) que en los últimos años han sido objeto de estudios de validación para determinar y confirmar su calidad psicométrica. Algunas de estas escalas incluyen, además de la evaluación clínica realizada por un profesional (neurólogo, fisioterapeuta, geriatra, etc.), una sección informada por el paciente, como sucede con la MDS-UPDRS.

Por otro lado, como consecuencia de la relevancia que ha adquirido en los últimos años la investigación de los síntomas no motores de la enfermedad, se han desarrollado diversos instrumentos para detectar y evaluar dichos síntomas, entre ellos, instrumentos comprehensivos de *screening* y evaluación, como el Cuestionario de Síntomas no Motores (*Non-motor Symptoms Questionnaire*, NMSQuest) y la Escala de Síntomas no Motores (*Non-motor Symptoms Scale*, NMSS) [24,25], respectivamente, y escalas dirigidas a evaluar ciertos dominios, como la serie de escalas SCOPA (Cognición, Sueño, Autonómica y Complicaciones Psiquiátricas) [26–29] o síntomas específicos, como la Escala de Fatiga de Parkinson (*Parkinson Fatigue Scale-16*, PFS-16) [30].

Tabla 1.2. Principales instrumentos utilizados para evaluar la enfermedad de Parkinson.

Dominio	Instrumentos
Escalas globales	- *Movement Disorders Society sponsored version of the Unified Parkinson's Disease Rating Scale*, MDS-UPDRS (Goetz et al, 2007) - Escala de Síntomas no motores. *Non-motor Symptoms Scale*, NMSS (Chaudhuri et al, 2007).
Gravedad	- Estadios de Hoehn y Yahr. *Hoehn & Yahr Staging*, HY (Hoehn y Yahr, 1967). - Índice de Impresión de Gravedad para enfermedad de Parkinson. *Clinical Impression of Severity Index for Parkinson's disease*, CISI-PD (Martínez-Martín et al., 2006)
Deterioro cognitivo	- *Mini-Mental State Examination*, MMSE (Folstein et al, 1975) - *Scales for Outcomes in Parkinson's disease*, SCOPA-Cognición (Marinus et al, 2003) - *Mattis Dementia Rating Scale*, DRS (Mattis, 1976) - *Montreal Cognitive Assessment*, MoCA (Nasreddine et al., 2005)
Depresión	- Escala de Depresión de Hamilton. *Hamilton Depression Scale*, Ham-D (Hamilton, 1960) - *Beck Depression Inventory*, BDI (Beck et al., 1961) - *Hospital Anxiety and Depression Scale*, HADS (Zigmond & Snaith, 1983)
Ansiedad	- Escala de Ansiedad de Hamilton. *Hamilton Anxiety Rating Scale*, HARS (Hamilton, 1959) - *Zung Self-rating Anxiety Scale*, SAS (Zung, 1971) - *Hospital Anxiety and Depression Scale*, HADS (Zigmond & Snaith, 1983)
Apatía	- Escala de Apatía. *Apathy Scale*, AS (Starkstein et al., 1992) - *Lille Apathy Rating Scale*, LARS (Sockeel et al., 2006)
Anhedonia	- *Snaith-Hamilton Pleasure Scale*, SHAPS (Snaith et al., 1995)
Síntomas psicóticos	- *Parkinson Psychiatric Rating Scale*, PPRS (Friedberg et al., 1998) - SCOPA-Complicaciones psiquiátricas. *SCOPA-Psychiatric Complications*, SCOPA-PC (Visser et al., 2007)
Síntomas autonómicos	- SCOPA-Autonómico. *SCOPA-Autonomic*, SCOPA-AUT (Visser et al., 2004)
Trastornos del sueño	- *Pittsburgh Sleep Quality Index*, PSQI (Buysse et al., 1989) - SCOPA-Sueño. *SCOPA-Sleep* (Marinus et al., 2003) - Escala de Somnolencia de Epworth. *Epworth Somnolence Scale*, ESS (Johns, 1991) - *Parkinson's Disease Sleep Scale*, PDSS (Chaudhuri et al., 2002)
Fatiga	- *Fatigue Impact Scale for Daily Use*, D-FIS (Fisk y Doble, 2002) - *Parkinson Fatigue Scale*, PFS-16 (Brown et al., 2005)
Dolor	- *McGill Pain Questionnaire* (Lee et al., 2006)

1.3. La medida de la salud y resultados comunicados por los pacientes en la enfermedad de Parkinson.

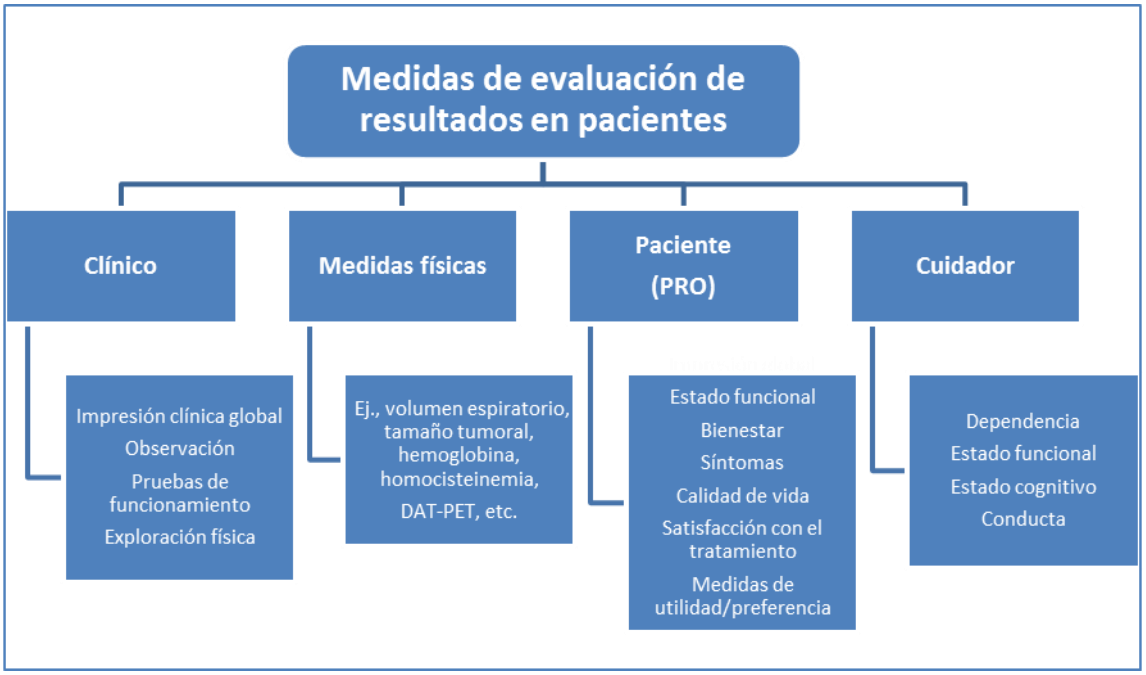
En los últimos años se ha introducido el concepto de "resultados comunicados por los pacientes" (en inglés "*patient reported outcomes"*, PRO) en la práctica y la investigación médica. Los resultados comunicados por los pacientes se pueden definir como "información sobre cualquier aspecto del estado de salud del paciente que proviene directamente de él, sin la interpretación de las respuestas por parte del clínico o cualquier otro profesional" [31]. Mediante los resultados comunicados por los pacientes se incorpora la perspectiva del paciente al proceso de evaluación y tratamiento de la enfermedad, generándose una información única, subjetiva y no directamente observable sobre aspectos personales, psicológicos y sociales relacionados con la enfermedad [32]. Esta información debe ser considerada junto con las variables clínicas, ya que ambos tipos de información son complementarios. En el caso de enfermedades crónicas e incapacitantes como la enfermedad de Parkinson, las medidas de resultados comunicados por los pacientes proporcionan indicadores que resultan de gran utilidad para evaluar la eficacia de intervenciones o modificaciones en el curso de la enfermedad, en ocasiones de manera más adecuada y exacta que los parámetros clínicos clásicos tales como la intensidad de los signos o los marcadores biológicos [33,34]. Por ello, las agencias reguladoras americana y europea (*Food and Drug Administration*, FDA, y *European Medicines Agency*, EMA) [31,35] han elaborado documentos que recogen la importancia de las medidas de resultados comunicados por los pacientes en ensayos clínicos, guiando su diseño y análisis de manera que se garantice su calidad [31,36,37]. Acquadro y colaboradores (2003) [32] destacan una serie de puntos en relación con el valor de las medidas de resultados comunicados por los pacientes:

- La perspectiva del paciente es un elemento clave en el diagnóstico y tratamiento médicos.
- Los resultados comunicados por los pacientes son indicadores complementarios de la actividad de la enfermedad y la eficacia del tratamiento.
- Las organizaciones profesionales reconocen la relevancia de los resultados informados por los pacientes en el diagnóstico y en el tratamiento, como se recoge en las correspondientes Guías de Práctica Clínica.

- En los ensayos clínicos, los resultados comunicados por los pacientes proporcionan una información muy importante para evaluar la eficacia del nuevo tratamiento.
- De acuerdo con su definición como instrumentos científicos, las medidas de resultados comunicados por los pacientes proporcionan datos precisos, fiables, válidos y reproducibles.
- La inclusión de medidas de resultados comunicados por los pacientes en ensayos clínicos es también apoyada por organizaciones profesionales, como se evidencia en las correspondientes guías desarrolladas por dichas organizaciones.
- Los resultados comunicados por los pacientes son esenciales para la práctica médica basada en la evidencia.
- Los datos de resultados comunicados por los pacientes procedentes de ensayos clínicos apoyan la práctica médica basada en la evidencia.

La relevancia de las medidas de resultados comunicados por los pacientes en ensayos clínicos y en la autorización de medicamentos queda patente en diversos estudios que muestran la expansión de su utilización bien como medidas primarias (*endpoint*) o complementarias de resultados [38–40]. Las medidas de resultados comunicados por los pacientes incluyen, pero no se limitan a, los siguientes aspectos: síntomas, percepción del cuidado, adherencia y satisfacción con el tratamiento, estado funcional y de salud, bienestar psicológico y calidad de vida [14,22]. Como se observa en la Figura 1.1, los datos para evaluación de resultados se pueden obtener a partir de valoraciones clínicas, pruebas de laboratorio, autoevaluaciones o mediante información a partir de terceros (cuidadores, familiares), lo que se conoce como "por delegación" o *by proxy*, una estrategia útil para obtener información cuando el paciente no puede expresarse. Sin embargo, este último tipo de información debe manejarse e interpretarse con cuidado ya que no siempre es equivalente a la proporcionada por el propio paciente [42].

Figura 1.1. Medidas de evaluación de resultados en pacientes, fuentes y ejemplos. Adaptada de Acquadro y colaboradores (2003).



A pesar de su utilidad, la utilización de medidas de resultados comunicados por los pacientes ha sido criticada debido a la naturaleza subjetiva de la información que aportan y a problemas metodológicos derivados de la escasez de datos de validación para algunos instrumentos. La elección de un instrumento de resultados comunicados por los pacientes para evaluar manifestaciones de la enfermedad de Parkinson debe guiarse por los mismos principios que se aplican a la medición clínica, tomando en consideración la adecuación del instrumento al objetivo del estudio y su calidad métrica.

1.4. Diseño y validación de escalas

Medir es un requisito imprescindible de la ciencia y puede ser definido como "la asignación de números, según normas, a objetos o eventos" [43]. Esto puede resultar sencillo cuando el objeto de medida es directamente observable y existe una unidad de referencia (por ejemplo, la altura, el peso o el tiempo). Sin embargo, determinados atributos o conceptos, que denominamos "constructos", son abstractos (por ejemplo, el dolor o la ansiedad), y no se pueden observar directamente. Su evaluación es compleja y peculiar: carecen de un patrón externo unitario, pueden ser interpretados de diversas

maneras y solo se pueden medir indirectamente, a través de aspectos observables relacionados con ellos. La metodología para evaluar constructos proviene de la psicología y las ciencias sociales, puesto que los atributos que caracterizan sus objetos de medida, como las emociones o el bienestar, pertenecen a dichas ciencias.

La medición de constructos se realiza mediante escalas y cuestionarios de evaluación que, debido a su facilidad de uso y coste reducido, son ampliamente utilizados en investigación clínica y en la práctica diaria. La elaboración de una escala de evaluación debe seguir una serie de principios tales como la inclusión de aspectos de interés real para los pacientes y de componentes estrechamente relacionados con dichos aspectos, la generación de puntuaciones que puedan tratarse estadísticamente y la facilidad de uso [44,45].

En Medicina, y por extensión en Neurología, las escalas de evaluación se utilizan con propósitos muy diversos: seguimiento del curso de la enfermedad, determinación de los factores asociados, comparación de estados de salud, evaluación de resultados de intervenciones y asignación de recursos. Son medidas con un considerable componente subjetivo debido a que se basan en la interpretación de lo observado y en la inferencia que conllevan al otorgar una puntuación concreta dentro del rango de opciones, por lo que conllevan un cierto grado de error. No obstante, existen diferentes formas de garantizar la calidad de la información que proporcionan, como un adecuado proceso de creación de la escala, la comprobación estandarizada de sus propiedades, instrucciones precisas para su aplicación y, si se requiere, el entrenamiento de los usuarios. Para muchos aspectos de interés en el contexto de la investigación y la práctica clínicas, las escalas de valoración constituyen el único sistema de medida disponible por lo que están justificados los esfuerzos para conseguir evaluaciones fiables y precisas.

El desarrollo de una escala sigue una metodología que consta de diversas fases: 1) especificación de los objetivos, propósito, formato y modo de aplicación de la escala; 2) selección de los componentes del instrumento (ítems, dominios), el marco temporal, el formato y redacción de los ítems y de las opciones de respuesta y el rango de puntuación; 3) realización de un estudio piloto para obtener la versión definitiva de la escala y 4) validación de la misma.

La definición clara y operativa del dominio de interés que se va a evaluar es un primer paso esencial a la hora de construir una escala de evaluación. La revisión de bibliografía, la consulta a expertos, pacientes y familiares, la realización de grupos de discusión (focales, paneles delphi) y el análisis cualitativo, son algunas de las técnicas a las que se recurre en esta fase para conseguir la definición del marco conceptual del constructo, sus componentes y dimensiones y de las relaciones entre ellos. De forma similar, y relacionado con el marco conceptual, es necesario determinar el propósito principal de la escala (discriminativo, evaluativo o predictivo), la población "diana" a la que se dirige, el formato y el modo de aplicación (p.ej., por teléfono, por entrevista, etc.). Una medida discriminativa permite distinguir entre grupos de individuos en un momento determinado, mientras que una evaluativa permite el seguimiento del estado de salud en el individuo o grupo de individuos a lo largo del tiempo. Las medidas predictivas son aquellas que anticipan una situación o pronóstico.

La generación y selección de ítems y componentes de la escala se deriva del proceso anterior. Los ítems deben reflejar áreas de interés para la población a la que va destinada la escala, por lo que dicha población debe implicarse durante la fase de generación y selección de ítems, sobre todo el el diseño de un instrumento de resultados comunicados por los pacientes [46].

El tercer paso, esencial para garantizar su calidad, es validarla siguiendo una metodología psicométrica o clinimétrica muy precisa. Validar una escala consiste en comprobar su calidad mediante el análisis estadístico de diferentes atributos psicométricos, siguiendo un conjunto de normas y estándares basado en métodos y teorías científicas de la medida de la salud. En este sentido, existen dos enfoques principales: la teoría clásica de la prueba (en inglés, *Classical Test Theory*) [44], que es la que se desarrollará en esta Tesis, y las teorías del rasgo latente (TRL), que incluyen a su vez dos enfoques: la teoría de la respuesta al ítem (TRI) y la metodología Rasch [47,48].

De acuerdo con los principios de validación de la teoría clásica de la prueba, el instrumento se pone a prueba en un conjunto amplio y representativo de la población a la que va destinada (muestra) y los datos se analizan para determinar una serie de propiedades: viabilidad y aceptabilidad, asunciones escalares, fiabilidad, validez, precisión, sensibilidad al cambio e interpretabilidad, aunque esta última no debe

considerarse una propiedad de la escala, sino una característica relevante de la misma [49,50]. A la hora de utilizar una escala es necesario tener en cuenta, además, la carga de aplicación del instrumento (conjunto de tiempo, esfuerzo y requerimientos necesarios para el uso de la escala), relacionada con la viabilidad, así como la posibilidad de que requiera adaptación transcultural, en el caso de que se vaya a utilizar en un contexto cultural o en una población diferente a la original (por ejemplo, en otro país). En función del propósito de la escala (discriminativo, evaluativo o predictivo), unos parámetros son más relevantes que otros. Así, en el caso de las escalas discriminativas, la fiabilidad y la precisión son atributos cruciales, mientras que para los instrumentos evaluativos y predictivos, es útil contar con datos de sensibilidad al cambio

En los últimos años se han desarrollado directrices para guiar el proceso de validación de escalas de medida de la salud y de resultados comunicados por los pacientes, tales como las del *Scientific Advisory Committee (SAC) of the Medical Outcomes Trust* [51], las iniciativas COSMIN (*Consensus-based Standards for the selection of health Measurement Instruments*) [50,52] y EMPRO (*Evaluating Measures of Patient Reported Outcomes*) [53], o las revisiones de Terwee y colaboradores [46] y Fitzpatrick y colaboradores [54]. Estas directrices definen los requisitos que debe cumplir un instrumento de medida del estado de salud o de resultados comunicados por los pacientes para garantizar su calidad. Sin embargo, dichas iniciativas difieren tanto en la terminología elegida para cada una de las propiedades métricas como en las definiciones, tal como se aprecia en la Tabla 1.3. Hay, no obstante, cierto consenso sobre los valores estándar que definen la calidad de cada una de las propiedades métricas (Tabla 1.4).

Tabla 1.3. Diferencias terminológicas entre los sistemas de clasificación de propiedades métricas de las escalas de evaluación.

MOS-SAC [51]	**COSMIN** [50,52]	**EMPRO** [53]	**Terwee y col.** [46]	**Fitzpatrick y col.** [54]
Modelo conceptual y de medida		Modelo conceptual y de medida	Validez de contenido	Adecuación
Fiabilidad: - Consistencia interna - Reproducibilidad	Fiabilidad: - Consistencia interna - Reproducibilidad - Error de medida	Fiabilidad: - Consistencia interna - Reproducibilidad	Consistencia interna Reproducibilidad: - Acuerdo - Fiabilidad	Fiabilidad: - Consistencia interna - Reproducibilidad
Validez: - De contenido - De constructo - De criterio	Validez: - De contenido y aparente - De constructo: estructural, comprobación de hipótesis - De criterio	Validez: - De contenido - De constructo - De criterio	Validez: - De constructo - De criterio	Validez: - Aparente y de contenido - De constructo - De criterio y predictiva
Sensibilidad al cambio	Sensibilidad al cambio	Sensibilidad al cambio	Sensibilidad al cambio	Sensibilidad al cambio
				Precisión
Interpretabilidad	Interpretabilidad	Interpretabilidad	Interpretabilidad	Interpretabilidad
			Efectos suelo y techo	Aceptabilidad
				Viabilidad
Carga: - de respuesta - administrativa		Carga: - de respuesta - administrativa		
Formas alternativas de administración		Formas alternativas de administración		
Adaptaciones culturales y lingüísticas y traducciones	Validez transcultural	Adaptaciones lingüísticas y culturales		

Tabla 1.4. Criterios o valores estándar para las diferentes propiedades psicométricas de acuerdo con la teoría clásica de la prueba.

PROPIEDAD Y DEFINICIÓN	CRITERIO	REFERENCIA
Viabilidad Hasta qué punto el instrumento puede ser utilizado en el contexto previsto.	**Datos perdidos** < 5 – 10%.	[55]
Aceptabilidad Hasta qué punto el instrumento es acceptable para la población a la que se dirige.	**Rango teórico – observado:** coincidentes. **Media - mediana:** coincidentes. **Efectos suelo y techo:** < 15%. **Asimetría:** entre –1 y +1.	[56–58]
Fiabilidad: grado en el que el instrumento está libre de error aleatorio.		
Consistencia interna Hasta qué punto los ítems miden el mismo constructo.	**Alfa de Cronbach:** > 0,70 (grupo). 0,90 – 0,95 (individual). **Correlación item-total:** r = 0,20 – 0,40. **Correlación inter-ítem:** r > 0,20 y < 0,75. **Coeficiente de homogeneidad:** > 0,30. Constructo amplio: > 0.15 o > 0.20. Constructo limitado: > 0.40 o > 0.50.	[55,56,59–61]
Reproducibilidad Estabilidad de las puntuaciones en el tiempo o en diferentes evaluadores.	**Fiabilidad inter-observador y test-retest:** Kappa (datos nominales/ordinales) > 0,60 o > 0,70. Coeficiente de correlación intraclase (datos continuos) > 0,70.	[46,62]
Validez: grado en el que el instrumento mide aquello que debe medir.		
Validez de contenido Grado en el que el instrumento recoge todos los componentes relevantes del constructo.	Opinión experta. Índice de Lynn de validez de contenido.	[63]
Validez de constructo (Comprobación de hipótesis) Grado en el que las puntuaciones del instrumento son consistentes con hipótesis referidas al constructo.	**Validez convergente**: r > 0,60 (0,50 – 0,70). **Validez divergente**: r < 0,30. **Validez interna**: r = 0,30 – 0,70. **Validez para grupos conocidos**: Diferencias significativas entre los grupos.	[35,54,64,65]
Validez de criterio Relación del instrumento con una medida de referencia (*gold standard*).	r > 0,60.	[52]

1.4.1. Modelo conceptual y de medida

Se trata de describir y definir el marco conceptual del constructo y la población a la que va dirigida la escala [46,51,53]. Aquí se incluye: la descripción clara y operativa del concepto que se va a medir, la definición del propósito de la escala (evaluativo, discriminativo o predictivo), las bases empíricas y conceptuales del contenido de los ítems y de la combinación de los mismos en una o más dimensiones, la implicación de la población diana en la definición del contenido de la escala y de los ítems, la descripción del nivel de medida elegido (ordinal, de intervalo, de razón) y la evidencia empírica que apoya dicha elección y el procedimiento para obtener las puntuaciones de la escala y de los dominios.

En la metodología COSMIN, este apartado se incluye en el epígrafe de "validez de contenido" [50,52], pero resulta evidente que es el paso inicial desde el punto de vista de desarrollo de la escala.

1.4.2. Viabilidad y aceptabilidad

La viabilidad es la aplicabilidad de la escala en el entorno previsto [54]. Para analizarla, se recurre a indicadores tales como el porcentaje de datos faltantes (que debe ser inferior al 10%) y el porcentaje de datos computables (que debe ser superior al 95%) [55].

La aceptabilidad indica hasta qué punto la escala recoge todo el espectro de intensidad del rasgo evaluado y si la distribución de las puntuaciones en la muestra es adecuada. Sus indicadores son la comparación del rango teórico de las puntuaciones con el observado en la realidad, la diferencia entre la media y la mediana, los efectos suelo y techo (es decir, el porcentaje de puntuaciones que se situan en el extremo inferior y superior de la escala de puntuación, que no debe ser mayor del 10 o el 15%, según los diferentes autores) [56] y la desviación de las puntuaciones, medida mediante los valores de asimetría, que deben oscilar entre -1 y +1, y curtosis (el valor 0 indica distribución normal) [57,58]. En algunos estudios también se emplean preguntas directas sobre la escala a los usuarios, sobre todo en el estudio piloto [54].

1.4.3. Fiabilidad

En la teoría clásica de la prueba, la puntuación observada de un individuo en una prueba es la suma de dos componentes: la puntuación verdadera y un componente aleatorio de error de medida. La fiabilidad nos informa de hasta qué punto una escala está libre de este error de medida. Consta de dos aspectos fundamentales: la consistencia interna y la estabilidad, o fiabilidad propiamente dicha [52]. La fiabilidad no es una propiedad fija del instrumento, sino que depende del contexto y de la población estudiada [54].

La consistencia interna se refiere al grado en que los ítems de una escala miden el mismo constructo. Se mide por lo general mediante índices como el alfa de Cronbach, para variables contínuas, o el Kuder-Richardson para ítems con respuestas dicotómicas. Para ambos índices se ha establecido el criterio de que el valor debe situarse entre 0,70 y 0,95 [51,55,56]. Otros métodos son la correlación ítem-total, la correlación inter-item y el coeficiente de homogeneidad de los ítems, que se obtiene al promediar los coeficientes de correlación inter-item [59–61].

La estabilidad incluye a la fiabilidad test-retest y a la fiabilidad inter-observador, que se calculan mediante índices como el simple porcentaje de acuerdo y, más apropiadamente, mediante coeficiente de correlación intra-clase (para variables continuas) y el coeficiente kappa (para variables categóricas u ordinales), que compensa el grado de acuerdo debido al azar [46,62].

El error de medida se define como "el error sistemático y aleatorio de la puntuación de un paciente que no es atribuible a cambio verdadero en el constructo que está siendo medido". Puede determinarse en situaciones test-retest o seguimiento longitudinal mediante el error estándar de la medida (EEM; en inglés, *standard error of measurement*, SEM), también considerado como medida de precisión o sensibilidad de la escala) y el menor cambio detectable (*smallest detectable change*) [52].

1.4.4. Validez

La validez determina si una escala realmente mide el constructo para el cual se diseñó. Existen varios tipos de validez, si bien las que se determinan con mayor frecuencia en los estudios son la validez de contenido y la validez de constructo. Con la primera se comprueba hasta qué punto los componentes de una escala (ítems, subescalas) son adecuados para medir el constructo, mediante juicio de expertos sobre la escala, el método de generación y selección de ítems, etc. Dicho juicio se puede materializar en una medida cuantitativa como el índice de Lynn [63].

La validez de constructo se puede definir como el grado en que un instrumento evalúa el constructo subyacente que se propone medir, de acuerdo con hipótesis derivadas teóricamente y referidas a dicho constructo. Se trata de comprobar diversas hipótesis que demuestren si la escala mide de forma válida el constructo que se supone que mide: por ejemplo, sobre las relaciones entre sus componentes, la asociación con otros instrumentos que miden el mismo constructo, o las diferencias en puntuaciones entre grupos relevantes (*known-groups validity*) [52].

Tradicionalmente, la validez de constructo se ha dividido en validez convergente, validez divergente o discriminante, validez discriminativa o para grupos conocidos y validez interna. La validez convergente establece la relación del instrumento con otros que miden el mismo constructo o constructos relacionados; mientras que la validez divergente o discriminante analiza el comportamiento de la escala frente a medidas de constructos diferentes. La validez discriminativa (o validez para grupos conocidos) se refiere a la capacidad de la escala para detectar diferencias entre grupos de sujetos que difieren en el constructo. La validez interna es el examen de la relación entre los distintos componentes (subescalas o dominios) de la escala. Los valores estándar fijados para comprobar la adecuación de la validez de constructo son los siguientes: para la validez convergente, se establece como mínimo una correlación de 0,60 entre las escalas analizadas (con rango entre 0,50 y 0,70, dependiendo de los diferentes autores); para la validez divergente o discriminante, el coeficiente de correlación debe ser menor de 0,30; para la validez interna, los coeficientes de correlación deben oscilar entre 0,30 y 0,70; mientras que para la validez discriminativa, deben detectarse diferencias estadísticamente significativas entre los grupos establecidos mediante la prueba de contraste de hipótesis apropiada [35,54,64,65].

La iniciativa COSMIN ha propuesto cambios en la terminología en relación con la validez. Así, distingue entre la validez de contenido y la validez facial o "aparente" (*face validity*), mientras que dentro de la validez de constructo (*hypotheses testing*) incluye la validez estructural y la validez trans-cultural (*cross-cultural validity*) [52].

Otro tipo de validez a considerar es la validez de criterio, entendida como la relación de una escala con una medida de referencia o *gold standard* (por ejemplo, una prueba de laboratorio) [52]. La dificultad de este tipo de validez radica en la definición y elección de la medida de referencia correcta. En ocasiones, se puede recurrir a una medida fisiológica o al juicio clínico basado en criterios diagnósticos. Sin embargo, en muchas ocasiones, como sucede con las medidas de resultados comunicados por los pacientes (por ejemplo, la calidad de vida), el *gold standard* puede simplemente no existir, razón que justifica de pleno la creación de un instrumento de medida. En ocasiones, una forma breve o modificada (por ejemplo, por un método de aplicación alternativa) se comparan frente al instrumento original del que provienen, considerando a éste como el *gold standard*.

1.4.5. Sensibilidad al cambio

La sensibilidad al cambio es un parámetro de gran importancia, pues de ella dependerán en parte los resultados de un seguimiento evolutivo o de un tratamiento. El término está íntimamente ligado a la "precisión" o "sensibilidad" del instrumento, que se define como "la capacidad de la medida para distinguir pequeñas diferencias". Cuanto más preciso (sensible) sea un instrumento, más probable es que pueda captar pequeños cambios, mientras que un instrumento poco sensible requerirá grandes cambios en el constructo para mostrar modificaciones en sus puntuaciones. La precisión de una escala se mide habitualmente mediante el error estándar de la medida (EEM), que se calcula multiplicando la desviación típica basal (DT) por $\sqrt{(1-r_{xx})}$, siendo r_{xx} el coeficiente de fiabilidad [51]. El EEM determina un umbral por encima del cual el cambio es considerado real (o mínimo cambio detectable), una vez eliminado el "ruido" o variabilidad en las puntuaciones que no se debe a un cambio real en el estado de salud.

La "sensibilidad al cambio" (*responsiveness*) se ha definido como la capacidad de una medida del estado de salud para captar cambios en dicho estado, aunque las características a considerar de dicho cambio (cualquier cambio, un cambio verdadero, clínico, global, mínimo, etc.) son imprecisas y dan lugar a multitud de definiciones diferentes. Por otra parte, hay autores que proponen que el cambio detectado debe ser clínicamente importante, lo que relaciona la sensibilidad al cambio con la interpretabilidad [46,51,66]. La sensibilidad al cambio depende de parámetros tales como la existencia de efectos suelo y techo, la fiabilidad y la precisión de la escala.

No hay consenso sobre la forma de medir la sensibilidad al cambio, aunque tradicionalmente se han utilizado pruebas de significación estadística (por ejemplo, pruebas t de Student, Wilcoxon o ANOVA para medidas repetidas), en las que se acepta o rechaza la hipótesis nula que establece que no hay diferencia entre las puntuaciones basal y final [67]. No obstante, las pruebas de significación estadística no informan de la dirección del cambio o de su importancia y además están influenciadas por el tamaño muestral.

Más adecuadas son las determinaciones basadas en la magnitud del cambio, como el cambio relativo (CR; en inglés, *relative change*), que se expresa como un porcentaje, y el tamaño del efecto (TE; *effect size*) y la respuesta medida estandarizada (RME; *standardized response mean*), que se calculan dividiendo la diferencia media final y basal de las puntuaciones por la desviación típica de la evaluación basal en el caso del tamaño del efecto, o por la desviación típica de la diferencia de puntuaciones, en el caso de la respuesta media estandarizada [66]. Estas medidas, TE y RME, transforman el cambio en las puntuaciones en unidades estandarizadas que pueden ser fácilmente interpretadas mediante puntos de referencia como los establecidos por Cohen: valores menores a 0,20 indicarían un efecto insignificante; entre 0,20 y 0,49, un efecto pequeño; entre 0,50 y 0,79, un efecto moderado; y mayores a 0,80 un efecto grande [68,69].

Otras medidas propuestas para evaluar la sensibilidad al cambio son la eficiencia relativa [70], la menor diferencia real (*smallest real difference*, SRD) [71], el índice de cambio fiable (*reliable change index*, RCI) [72,73] y el índice de Guyatt (en inglés, *Guyatt's responsiveness statistics*) [74,75].

1.5. Interpretación de resultados (interpretabilidad)

Una vez que se ha determinado que existe un cambio real en la medida analizada, es necesario estudiar la importancia del cambio, proceso relacionado con la interpretabilidad [51], la cual se define como "el grado en que se puede asignar un significado fácilmente comprensible a las puntuaciones cuantitativas de un instrumento" [46]. Se trata de un concepto que ha cobrado gran relevancia en los últimos años en consonancia con los esfuerzos para desarrollar directrices de validación de escalas, si bien la interpretabilidad no es una propiedad psicométrica propiamente dicha, sino un aspecto relacionado con la utilidad y la sensibilidad al cambio de la escala [50].

La interpretación de puntuaciones basadas en una única medición (estudios transversales) se realiza mediante la comparación de las puntuaciones medias de la muestra con normas poblacionales o la comparación de puntuaciones entre grupos de pacientes que, por estudios anteriores, se sabe que presentan diferencias en el constructo que se está midiendo, mediante la correlación con otras medidas fácilmente interpretables, etc.

En estudios longitudinales, las técnicas de interpretación del cambio se basan principalmente en el cálculo del mínimo cambio importante (MCI; en inglés, *minimally important change, MIC*) o la mínima diferencia importante (MDI; *minimally important difference,MID*). La metodología para dicho cálculo se puede clasificar en dos grandes aproximaciones: técnicas basadas en un criterio externo o anclaje (*anchor*) y técnicas basadas en la distribución de puntuaciones [69,76]. A estas técnicas, nuestro grupo ha añadido la "magnitud del cambio", mostrada a través de diversos parámetros, que permite un juicio intuitivo y directo de la importancia de las modificaciones entre el período basal y el seguimiento.

1.5.1. Técnicas basadas en un criterio externo

Las técnicas basadas en un criterio externo (anclaje) comparan el cambio en un concepto medido por una escala y el cambio en el mismo concepto o uno relacionado medido mediante otra escala o una medida que actúa como referencia. Uno de los tipos más habituales de criterio externo son las escalas de auto-evaluación del cambio [77].

De esta forma, el paciente debe responder a una pregunta, denominada "pregunta de transición", que se formula habitualmente como "¿hasta qué punto cree que ha cambiado su enfermedad, síntoma, etc. (el constructo de interés)?" entre dos momentos temporales relevantes (por ejemplo, antes y después de la intervención, o al comienzo y al final del estudio). Las posibles respuestas, en general, oscilan entre "mucho peor" y "mucho mejor" y no deben tener más de 7 opciones (tres de gradación de mejoría, una para la ausencia de cambio y tres de gradación de empeoramiento) para garantizar una clara distinción entre las mismas [76]. Dichas respuestas permiten clasificar a la muestra en grupos de pacientes que han experimentado un cambio (mejorando o empeorando) y pacientes estables, y calcular las diferencias en las medias de las escalas de evaluación en los pacientes que declaran haber mejorado (o empeorado) mínimamente para determinar cuál es el MCI o MDI de dichas escalas. Además de las escalas de cambio, también pueden utilizarse como criterio externo de impacto parámetros tales como la comparación con eventos vitales adversos (ejemplo, pérdida del empleo, grado de discapacidad, etc.), el resultado de pruebas clínicas, el uso de servicios, costes, etc.

Al usar técnicas basadas en un criterio externo, es necesario tener en cuenta varios aspectos. Así, el anclaje o criterio externo elegido debe ser adecuado, estrechamente relacionado con el objetivo de interés, fácilmente interpretable, con opciones de respuesta definidos con precisión, y con un coeficiente de correlación de al menos 0,30-0,35 con el instrumento de medida [34]. Además, la pregunta sobre el cambio puede verse afectada por diversos factores que influyen en su variabilidad, como la selección y composición de la muestra y su situación basal, así como los sesgos relativos al recuerdo, por lo que se recomienda recalcular el MCI en cada nuevo estudio así como el uso de varios criterios a la vez para ofrecer un rango de valores que, previsiblemente, incluya al MCI real (triangulación) [34,78].

1.5.2. Técnicas basadas en la distribución de las puntuaciones

Las técnicas basadas en la distribución de puntuaciones estiman el cambio de manera estandarizada a partir de métodos estadísticos que relacionan la magnitud del efecto con alguna medida de variabilidad [76,79]. La medida de la variabilidad puede ser inter- (por ejemplo, la desviación típica basal de los pacientes) o intra-sujetos (por ejemplo, la desviación típica del cambio que los pacientes experimentan durante un

estudio) [79] y justifican el empleo del tamaño del efecto y la respuesta media estandarizada, respectivamente, para la interpretación de resultados [69].

Diversos estudios han comprobado que el MCI determinado por técnicas de anclaje se aproxima a umbrales de cálculo sencillo (½ desviación típica basal, DT_{basal} o el 5%-10% de la puntuación total de la escala) [80–82], o algo más complejo (1 EEM), por lo que estos valores han sido propuestos como formas de estimación del MCI con base estadística [34,83–85]. Además, estos niveles de cambio (p.ej., 1 EEM o ½ DT_{basal}) pueden ser utilizados como la referencia para el cálculo del número necesario de pacientes a tratar (NNT) para obtener una mejoría (o empeoramiento) [79].

La principal ventaja de las técnicas basadas en la distribución de las puntuaciones es que ofrecen estandarización y permiten la interpretación del cambio cuando no existe un criterio externo para hacerlo. Sin embargo, precisamente por no estar relacionadas con una medida externa de cambio (fundamentalmente con una derivada de los propios pacientes) la interpretación es arbitraria, como ocurre con los puntos de corte establecidos por Cohen para el tamaño del efecto [68,69]. Por ello, para la interpretación de la importancia del cambio se recomienda el uso de técnicas basadas en un criterio externo (anclaje), utilizando las basadas en la distribución como medidas complementarias de interpretabilidad [34,76].

Típicamente, los métodos basados en distribución arrojan valores diferentes entre sí, por lo que proponemos – al igual que para las técnicas de anclaje – realizar una triangulación entre diferentes métodos, asumiendo que el verdadero valor del MCI se encuentra entre ellos.

1.5.3. Técnicas basadas en la magnitud del cambio

La magnitud del efecto de la progresión de una enfermedad o del resultado de un tratamiento se puede expresar de varias formas, como la diferencia de puntuaciones de un paciente o de un grupo de pacientes antes y después del tratamiento o del período de observación, o entre el grupo de control y el grupo experimental, el porcentaje de cambio [86] y la tasa anual de cambio [87,88].

La magnitud del cambio aporta una apreciación subjetiva de su importancia. Por ejemplo, una mejoría de 14 puntos (40%) en una escala de puntuación máxima 35

sugiere un cambio importante. No obstante, sin unos umbrales o puntos de corte definidos los límites entre grados o niveles de cambio son inciertos (en el ejemplo anterior, un cambio de 3 puntos ¿es importante o trivial?). Además, es necesario tener en cuenta la situación basal, pues la modificación experimentada en poblaciones con diferentes rangos de intensidad del constructo puede ser importante en los extremos y anodina en el centro de la escala.

OBJETIVOS

2. Objetivos

2.1. Objetivos generales

1) Validación de diversos instrumentos de medida de la salud (escalas clínicas y resultados comunicados por los pacientes) para enfermedad de Parkinson

2) Interpretación del cambio en una cohorte de pacientes con enfermedad de Parkinson utilizando dichas medidas.

2.2. Objetivos específicos

Siguiendo la metodología de validación de medidas de la salud basada en la teoría clásica de la prueba, se han realizado varios estudios (a nivel nacional e internacional) de validación e interpretabilidad de diversas escalas clínicas y de resultados comunicados por los pacientes para uso en enfermedad de Parkinson con el siguiente

Objetivo 1.

En estudios transversales, **analizar las propiedades psicométricas**, a partir de un estudio transversal, de escalas clínicas y de resultados comunicados por los pacientes desarrolladas por nuestro grupo o validadas independientemente por primera vez tras su desarrollo por otros grupos.

Dado que la enfermedad de Parkinson es progresiva, la utilización de instrumentos evaluativos en diversos momentos temporales permite determinar el cambio en el estado de salud del paciente (en este caso, la tendencia hacia el empeoramiento) que se refleja en la modificación de las puntuaciones. El segundo objetivo se centra en mostrar la validez de las escalas para detectar un cambio:

Objetivo 2.

En el estudio longitudinal con un seguimiento de tres años con las escalas validadas en el Estudio Longitudinal de Enfermedad de Parkinson (ELEP), **analizar la sensibilidad al cambio** (validez longitudinal) [66], de dichas escalas.

Con la intención de informar sobre la importancia que dicha modificación puede tener sobre el estado de salud del paciente, se ha abordado la interpretabilidad del cambio en las puntuaciones. El tercer objetivo se refiere a la aplicación de técnicas estadísticas sobre las puntuaciones basal y de seguimiento para tratar de establecer si el cambio fue probablemente importante o no:

Objetivo 3.

Estimar la **importancia del cambio**, mediante determinaciones de su magnitud y métodos basados en distribución en las escalas mencionadas.

MATERIAL Y MÉTODOS

3. Material y métodos

3.1. Diseño

Los datos de esta Tesis se obtuvieron a partir de tres estudios multicéntricos diferentes:

1. Estudio longitudinal de la enfermedad de Parkinson (ELEP), realizado a nivel nacional [89- 93].
2. Estudio de validación internacional de la escala de síntomas no motores (*Non-motor Symptoms Scale*, NMSS) [94].
3. Estudio de validación internacional de la versión en español de la escala *Movement Disorders Society-sponsored version of the Unified Parkinson's Disease Rating Scale* (MDS-UPDRS) [95].

3.2. Participantes

En los tres estudios se incluyeron pacientes mayores de 30 años diagnosticados de enfermedad de Parkinson por un neurólogo experto de acuerdo con los criterios del Banco de Cerebros de la United Kingdom Parkinson's Disease Society [1]. Los participantes eran pacientes de los departamentos de Neurología o unidades de Trastornos del Movimiento de diversos hospitales y fueron evaluados en el transcurso de visitas clínicas rutinarias. En el caso del estudio ELEP, se seleccionaron los pacientes por bloques en función del sexo, la edad de inicio (antes o después de los 60 años) y la duración de la enfermedad (menor o igual a 5 años o más de 5 años) para garantizar la presencia de un número similar de pacientes en todos los niveles de gravedad de la enfermedad [93]; en el resto fueron pacientes consecutivos.

Los criterios de exclusión de los pacientes fueron la incapacidad para leer, entender o responder cuestionarios y la presencia de cualquier trastorno, situación o

estado de salud que impidiera o interfiriera la evaluación de la enfermedad de Parkinson. Los pacientes con demencia fueron excluídos en todos los estudios.

La muestra del estudio ELEP estuvo compuesta por 387 pacientes en la evaluación basal y 228 al final del seguimiento. La muestra del estudio NMSS fue de 411 pacientes mientras que la del estudio MDS-UPDRS fue de 435 participantes.

Los tres estudios recibieron la aprobación del Comité de Ética de la Investigación y de Bienestar Animal (CEIyBA) del Instituto de Salud Carlos III y de los comités equivalentes en cada uno de los centros participantes. Todos los pacientes incluidos en los estudios firmaron el correspondiente consentimiento informado.

3.3. Evaluaciones

En los tres estudios citados se validaron las siguientes escalas clínicas y medidas de resultados comunicados por los pacientes:

3.3.1. Medidas clínicas

1. *Clinical Impression of Severity Index – Parkinson's Disease* (CISI-PD) [89,96]: se trata de una escala de impresión clínica global de la enfermedad de Parkinson que consta de cuatro dominios (síntomas motores, discapacidad, complicaciones motoras y estado cognitivo). Cada dominio se puntúa en un rango de 0 (normal) a 6 (muy grave). La puntuación total, que oscila entre 0 y 24 puntos, es el resultado de sumar los cuatro dominios. A mayor puntuación, mayor gravedad global de la enfermedad.
2. *Non-motor Symptoms Scale* (NMSS) [25,94]: es la primera escala global de evaluación de los síntomas no motores desarrollada para enfermedad de Parkinson. Consta de 30 ítems, agrupados en nueve dominios: cardiovascular (dos ítems), sueño/fatiga (cuatro ítems), estado de ánimo/apatía (seis ítems), problemas perceptuales/alucinaciones (tres ítems), atención/memoria (tres ítems), sistema gastrointestinal (tres ítems), urinario (tres ítems), función sexual (dos ítems) y miscelánea (cuatro ítems). Cada ítem tiene dos puntuaciones:

gravedad (de 0 a 3) y frecuencia (de 1 a 4), para poder evaluar síntomas que son graves pero infrecuentes o que son menos graves pero persistentes. La puntuación total de cada ítem se obtiene por multiplicación de las puntuaciones de gravedad por frecuencia y se obtienen, por suma, las puntuaciones totales por dominio y total de la escala (ésta oscila de 0 a 360 puntos), que representa la carga sintomática. Una mayor puntuación representa mayor afectación por los síntomas no motores.

3. *Modified Parkinson Psychosis Rating Scale* (mPPRS) [92,97]: se trata de una versión revisada y ampliada de la *Parkinson Psychosis Rating Scale*, el primer instrumento específico para evaluar síntomas psicóticos en enfermedad de Parkinson. La versión modificada consta de 6 ítems que evalúan la presencia y gravedad de alucinaciones, delirios y errores en la identificación de personas, ideación paranoide, trastornos del sueño, confusión y preocupación sexual. Los ítems se puntúan en una escala tipo Likert de 0 (ausente) a 3 (grave). A diferencia de la escala original, el ítem sobre alucinaciones de la mPPRS incluye todo tipo de alucinaciones, y la puntuación del ítem 2 (delirios y errores en la identificación de personas) se cambió para basarla en la intensidad, como el resto de ítems, y no en la frecuencia como figura en la versión original. En la escala mPPRS, una mayor puntuación indica una mayor gravedad de la síntomatología psicótica
4. *Movement Disorders Society sponsored version of the Unified Parkinson's Disease Rating Scale* (MDS-UPDRS) [16,94]: es la versión revisada y mejorada de la UPDRS, la escala de referencia para las agencias reguladoras europea y estadounidense, y la más ampliamente utilizada para evaluar la enfermedad de Parkinson [15,98]. Está compuesta por cuatro secciones: Parte I, aspectos no motores de las experiencias de la vida diaria (nM-EVD), con seis ítems administrados por el evaluador y siete ítems autoadministrados; Parte II, aspectos motores de las experiencias de la vida diaria (M-EVD), con 13 ítems auto-administrados; Parte III, exploración motora, que contiene 18 ítems evaluados por un clínico; y Parte IV, complicaciones motoras, con seis ítems también hetero-administrados. Los ítems se puntúan en una escala de 0 (normal) a 4 (grave) y se obtiene una puntuación total para cada una de las secciones.

3.3.2. Medidas de resultados comunicados por los pacientes

1. *Hospital Anxiety and Depression Scale* (HADS) [90,99]: esta escala se desarrolló para pacientes ambulatorios de las consultas hopitalarias y durante años se ha utilizado en pacientes con enfermedad de Parkinson debido a la ausencia de preguntas relativas a síntomas físicos de la ansiedad y la depresión. La escala HADS consta de siete ítems que evalúan ansiedad y siete para depresión, que se puntúan en una escala de 0 (sin problemas) a 3 (problema grave). Los ítems se suman para obtener una puntuacion para ansiedad, otra para depresión y puede calcularse otra, sumatorio de las anteriores, para trastorno del estado de ánimo en general (aunque existe debate sobre la conveniencia o no de calcular esta última). Se ha propuesto un punto de corte de 11 o más en cada subescala (ansiedad o depresión) para el diagnóstico de caso clínico.
2. *Scales for Outcomes in Parkinson's Disease – Autonomic* (SCOPA-AUT) [28,91]: es la primera escala específicamente desarrollada para evaluar síntomas autonómicos en la enfermedad de Parkinson. Está compuesta por 25 ítems que se agrupan en los siguientes dominios: gastrointestinal, urinario, cardiovascular, termorregulatorio, pupilomotor y sexual. Los ítems se puntúan en una escala de 0 (nunca) a 3 (a menudo), con una puntuacion total máxima de 69. Los dominios urinario y sexual incluyen la opciones de respuesta "uso catéter" y "no aplicable", respectivamente. A mayor puntuación, mayor disfunción autonómica.

Las escalas NMSS, mPPRS, MDS-UPDRS y SCOPA-AUT fueron traducidas y adaptadas transculturalmente (adaptación idiomática y conceptual) por el método de traducción y retrotraducción por personas bilingües, con participación de profesionales sanitarios, pacientes, lingüistas y expertos en el desarrollo y uso de escalas [100]. Las versiones traducidas y adaptadas fueron consideradas por los autores de las versiones originales como equivalentes a éstas en cada idioma y marco cultural de los países participantes en los referidos estudios.

3.4. Análisis de datos

Las bases de datos de los correspondientes estudios fueron diseñadas y centralizadas en el Centro Nacional de Epidemiología (Instituto de Salud Carlos III) con una fuerte implicación de la doctoranda en su creación, mantenimiento, supervisión y explotación.

En todos los estudios mencionados se aplicó a las escalas descritas en el apartado anterior la metodología de validación sistematizada y los estándares obtenidos de una amplia revisión bibliográfica y conceptual sobre el tema llevada a cabo en años precedentes por nuestro grupo. Para responder al **Objetivo 1**, en cada escala se analizaron los siguientes atributos psicométricos:

1. Viabilidad y aceptabilidad: porcentaje de datos perdidos y computables, distribución de las puntuaciones (medidas de tendencia central y dispersión), efectos suelo y techo, y asimetría.
2. Fiabilidad: consistencia interna mediante alfa de Cronbach, correlación ítem-total corregida. índice de homogeneidad de los ítems y estabilidad temporal (test-retest).
3. Validez de constructo por comprobación de hipótesis (validez convergente, validez para grupos conocidos y validez interna o estructural).
4. Precisión: error estándar de la medida (EEM).

Para el Objetivo 2, basado en el estudio longitudinal, se analizó la sensibilidad al cambio de aquellas escalas en las que se contaba con datos de seguimiento (CISI-PD, mPPRS, SCOPA-AUT y HADS), mediante los siguientes estadísticos: tamaño del efecto y respuesta media estandarizada (que también se utilizan para interpretación del cambio) y correlación del cambio en las puntuaciones con otras medidas.

Para el Objetivo 3, interpretación del cambio en las puntuaciones, se aplicaron los siguientes análisis:

1. Determinación de la magnitud del cambio (diferencia basal-final, cambio relativo, tasa anual de cambio), e interpretación del tamaño del efecto.
2. Identificación de umbrales derivados de la distribución de los datos (EEM, ½ DT_{basal}) o de la magnitud del cambio (10% del total teórico de la escala), seguida de triangulación con dichos valores umbral.
3. Descripción del cambio de puntuación (porcentaje de pacientes que modifican su estado a partir de un umbral específico y dirección del cambio) y determinación del número necesario de pacientes a observar para detectar un cambio en un caso en la tendencia de interés predominante (empeoramiento, en una enfermedad progresiva como es la enfermedad de Parkinson).

El análisis de datos se efectuó mediante los programas Stata/IC 13.0 e IBM SPSS Statistics 22.

RESULTADOS

4. Resultados

Los resultados se presentan en dos grandes bloques: el primer bloque resume los resultados de los estudios de validación de las escalas analizadas, mientras que en el segundo bloque se presentan los resultados del estudio de sensibilidad al cambio e interpretabilidad de varias de las escalas analizadas, realizado específicamente para esta Tesis.

4.1. Resultados de los estudios de validación de medidas clínicas en enfermedad de Parkinson

Como resultado de la aplicación de la sistemática de validación desarrollada y siguiendo la terminología propuesta por la iniciativa COSMIN [52], a continuación se presenta una síntesis descriptiva de las propiedades métricas analizadas para cada escala (Tabla 4.1). Se acompaña de una valoración del grado de cumplimiento de los criterios o valores estándar de dichas propiedades psicométricas, o calificación global: satisfactoria, si se cumplen todos los criterios; aceptable, si alguno de los atributos analizados no alcanza el valor criterio establecido; e indeterminada, si alguno de los atributos no se ha analizado (Tablas 4.2, 4.3, 4.4 y 4.5).

Tabla 4.1. Propiedades métricas estudiadas en cada una de las medidas (siguiendo la terminología COSMIN) [52]**.**

PROPIEDAD	CISI-PD	HADS	NMSS	SCOPA-AUT	mPPRS	MDS-UPDRS
Viabilidad/aceptabilidad	✓	✓	✓	✓	✓	✓
Fiabilidad						
Consistencia interna	✓	✓	✓	✓	✓	✓
Reproducibilidad: estabilidad temporal (fiabilidad test-retest)	✓		✓			✓
Validez						
Validez de contenido					✓	
Validez de constructo:						
- Validez convergente/grupos conocidos (comprobación de hipótesis, *hypotheses testing*)	✓	✓	✓	✓	✓	✓
- Validez interna/estructural	✓	✓	✓	✓	✓	✓
- Validez trans-cultural (*cross-cultural validity*)	✓			✓	✓	✓
Precisión	✓	✓	✓	✓	✓	✓

CISI-PD: Clinical Impression of Severity Index- Parkinson's Disease; HADS: Hospital Anxiety and Depression Scale; NMSS: Non-motor Symptoms Scale; SCOPA-AUT: Scales for Outcomes in Parkinson's Disease-Autonomic; mPPRS: modified Parkinson's Psychosis Rating Scale; MDS-UPDRS: Movement Disorders Society sponsored version of the Unified Parkinson's Disease Rating Scale.

4.1.1. Viabilidad y aceptabilidad

En la Tabla 4.2 se muestran los resultados del análisis de la viabilidad y la aceptabilidad de los instrumentos analizados. Todos mostraron un porcentaje de datos computable cercano al 100%, si bien dicho porcentaje fue menor en la SCOPA-AUT. En general, las escalas no mostraron efectos suelo y techo en sus puntuaciones totales o subescalas, exceptuando la presencia de efecto suelo en la MDS-UPDRS Parte IV y en la mPPRS. La asimetría se mantuvo dentro de los límites aceptados para todas las escalas, aunque se detectaron elevaciones marginales en la mPPRS y NMSS.

4.1.2. Fiabilidad

En relación con la fiabilidad, se analizaron los coeficientes alfa de Cronbach, correlación ítem-total corregida y homogeneidad de los ítems para la consistencia interna, y la fiabilidad test-retest para la reproducibilidad o estabilidad temporal (Tabla 4.3). Se observa que las escalas, en general, muestran una consistencia interna satisfactoria, aunque la escala mPPRS obtuvo un alfa de Cronbach 0,66, discretamente inferior al valor estándar (0,70).

En cuanto al coeficiente de correlación ítem-total corregido, en la SCOPA-AUT y en la MDS-UPDRS se detectaron ítems con coeficientes inferiores a 0,20. El índice de homogeneidad de los ítems fue inferior a 0,20 también en algunas subescalas de la NMSS, SCOPA-AUT y MDS-UPDRS.

Tabla 4.2. Principales atributos relativos a la viabilidad y aceptabilidad de las escalas analizadas-

ESCALA	Datos computables (%)	Puntuación total[a]	Efecto suelo (%)	Efecto techo (%)	Asimetría	Calificación global
CISI-PD	n.a.	8,30±4,51 (0-21)	0,33	0,49	0,50	**Aceptable**
HADS	A: 99,7 D: 99,5	A: 7,18±4,22 (0-20) D: 5,93±4,18 (0-21)	A: 3,1 D: 4,2	A: 0,3 D: 0,5	A: 0,52 D: 0,99	**Satisfactoria**
NMSS	99,5	57,1±44,0 (0-233)	0,5	0,2	1,2	**Satisfactoria**
SCOPA-AUT	96,9	20,41±11,1 (0-61)	0,27	0,27	0,40	**Satisfactoria**
mPPRS	99,7	1,26±1,83 (0-10)	46,65	0	2,6	**Aceptable**
MDS-UPDRS	I: 99,71 II: 99,08 III: 99,54 IV: 100	I: 11,27±6,97 (0-40) II: 14,78±9,56 (1-44) III: 32,46±16,30 (3-92) IV: 4,31±4,23 (0-21)	I: 0,69 II: 2,07 III: 0,23 IV: 30,34	I: 0,23 II: 0,46 III: 0,23 IV:0,23	I: 0,96 II: 1,01 III: 0,75 IV: 0,92	**Satisfactoria**

A: subescala Ansiedad; D: subescala Depresión; I: Aspectos no motores de las experiencias de la vida diaria (nM-EVD); II: Aspectos motores de las experiencias de la vida diaria (M-EVD); III: Exploración motora; IV: Complicaciones motoras; n.a.: no analizado.

[a] Media ± desviación típica (rango)

La fiabilidad test-retest se estudió en las escalas CISI-PD, NMSS y MDS-UPDRS. El coeficiente de correlación intraclase (CCI) del CISI-PD fue de 0,84, si bien el índice kappa del ítem exploración motora fue de 0,59. En el caso de la NMSS, el CCI fue de 0,90 para la puntuación total, y de entre 0,67 (función sexual) y 0,91 (estado de ánimo/apatía) para los dominios. El CCI de cada una de las secciones de la MDS-UPDRS fue superior a 0,90.

Tabla 4.3. Atributos relativos a la fiabilidad de las escalas analizadas.

ESCALA	Consistencia interna			Fiabilidad test-retest	Calificación global
	Alfa de Cronbach	Correlación ítem-total	Índice de homogeneidad		
CISI-PD	0,81	0,50-0,78	0,54	0,84	**Satisfactoria**
HADS[a]	0,81-0,83	0,39-0,72	0,38-0,40	n.a.	**Satisfactoria/ indeterminada**
NMSS[a]	0,44-0,85	0,20-0,73	0,16-0,54	0,67-0,91	**Aceptable**
SCOPA-AUT[a]	0,64-0,95	0,14-0,67	0,24-0,39	n.a.	**Satisfactoria/ indeterminada**
mPPRS	0,66	0,37-0,55	0,23	n.a.	**Aceptable/ indeterminada**
MDS-UPDRS[a]	0,79-0,93	0,18-0,79	0,20-0,49	0,92-0,97	**Satisfactoria**

n.a.: no analizado.

[a] Se muestran los datos para las subescalas.

4.1.3. Validez

En la Tabla 4.4 aparecen los datos relativos al análisis de comprobación de hipótesis (validez convergente, de grupos conocidos y validez interna o estructural) de las escalas.

Tabla 4.4. Atributos relativos a la validez de constructo de las escalas analizadas.

ESCALA	Comprobación de hipótesis (validez de constructo)			Calificación global
	Validez convergente[a]	Validez para grupos conocidos	Validez interna/ estructural	
CISI-PD	0,79 (HY) 0,85 (SCOPA-M) -0,46 (SCOPA-COG)	Gravedad (HY); duración de la enfermedad (p<0,001)	AFC: un factor (67% de la varianza)	**Satisfactoria**
HADS[b]	0,57-0,67 (SCOPA-PS) -0,48 - -0,56 (Índice EQ-5D)	Gravedad (HY, CISI-PD), duración de la enfermedad (p<0,001)	AFE: 2 factores (49,8% de la varianza)	**Satisfactoria**
NMSS	0,64 (SCOPA-AUT) 0,70 (PDQ-39) 0,57 (Índice EQ-5D)	Gravedad (HY, CISI-PD), duración de la enfermedad (p<0,001)	Correlaciones interdominio[a]: 0,06-0,42	**Satisfactoria**
SCOPA-AUT	0,55 (SCOPA-PS) 0,49 (SCOPA-M) 0,47 (HADS-D) -0,49 (Índice EQ-5D)	Edad, gravedad (HY y CISI-PD), duración de la enfermedad (p<0,001)	AFE: 8-9 factores, dependiendo del sexo (65-68% de la varianza). Correlaciones interdominio[a]: 0.17-0,47	**Aceptable**
mPPRS	0,56 (CISI-PD, Alucinaciones)	Gravedad (HY) (p<0,001)	AFE: 2 factores (58,5% de la varianza). AP: un factor.	**Satisfactoria**
MDS-UPDRS[b]	0,42-0,58 (HY) 0,42-0,81 (NMSS) 0,60-0,73 (CISI-PD) 0,47-0,72 (PDQ-8)	Edad, gravedad (HY) y duración de la enfermedad (p<0,001)	Correlaciones interdominio[a]: 0,41-0,61.	**Satisfactoria**

[a] Coeficientes de correlación de Spearman.

[b] Se muestran los datos para las subescalas.

HY: Estadios de Hoehn y Yahr; SCOPA: Scales for Outcomes in Parkinson's Disease; M: Motor; COG: Cognición; AFC: análisis factorial confirmatorio; PS: Psicosocial; AFE: análisis factorial exploratorio; AUT: Autonómico; D: depresión; AP: análisis paralelo.

La validez de contenido solo se analizó de manera formal en el caso de la mPPRS, mediante el índice de Lynn. Los ítems de esta escala se consideraron relevantes o altamente relevantes en el 80,6% de las evaluaciones realizadas por un panel de expertos.

El análisis de los coeficientes de correlación de las escalas analizadas con otros instrumentos relacionados con el correspondiente constructo evaluado arrojó valores superiores a 0,60 en la mayor parte de los casos. Hay que destacar las altas correlaciones observadas entre el CISI-PD e instrumentos de evaluación de manifestaciones motoras de la enfermedad (Hoehn y Yahr, HY, y SCOPA-M); la NMSS y una medida de calidad de vida específica para enfermedad de Parkinson, la escala *Parkinson's Disease Questionnaire*, 39-item (PDQ-39); y las distintas secciones de la MDS-UPDRS y los correspondientes instrumentos de evaluación de manifestaciones similares. Así, por ejemplo, la Parte I de la MDS-UPDRS, síntomas no motores, obtuvo un coeficiente de correlación de 0,81 con la NMSS, mientras que la Parte II, aspectos motores, mostró altas correlaciones con la Escala Rápida de Evaluación de la Discapacidad (*Rapid Assessment of Disability Scale*, RADS) (r=0,80) y con el ítem de discapacidad del CISI-PD (r=0,70).

En relación con la validez para grupos conocidos, todas las escalas obtuvieron puntuaciones significativamente diferentes en los pacientes agrupados según los niveles de gravedad de la enfermedad, definidos mediante HY o CISI-PD. Las escalas, excepto en el caso de la mPPRS, también distinguieron entre grupos de pacientes en función de la duración de la enfermedad. Los pacientes de más edad obtuvieron puntuaciones significativamente más altas en las escalas SCOPA-AUT y MDS-UPDRS. No se encontraron diferencias por sexo en ninguna de las escalas.

Por último, la validez estructural o interna se estudió mediante las correlaciones inter-dominio en las escalas multidimensionales (NMSS, SCOPA-AUT y MDS-UPDRS), y mediante análisis factorial exploratorio (HADS, SCOPA-AUT y mPPRS), confirmatorio (CISI-PD) o paralelo (mPPRS). Utilizando la primera técnica, en el caso de la NMSS y la SCOPA-AUT se observaron dominios con coeficientes de correlación inferiores a 0,30. Por su parte, la MDS-UPDRS obtuvo coeficientes de correlación de entre 0,41 y 0,61 entre las distintas secciones que la componen. Utilizando el análisis factorial exploratorio, se detectaron dos factores en las escalas HADS y mPPRS. Sin

embargo, para esta última un análisis paralelo (*parallel analysis*) [101] solo identificó un factor. En la SCOPA-AUT se detectaron ocho factores para los hombres y nueve factores para las mujeres. El análisis factorial confirmatorio ratificó la estructura unidimensional del CISI-PD.

4.1.4. Precisión

El error estándar de la medida se calculó en todas las escalas analizadas (EEM= ½ DT_{basal} * √(1-r_{xx}), siendo r_{xx} el coeficiente de fiabilidad). Para las escalas en que no se disponía de retest se utilizó como índice de fiabilidad el coeficiente alfa de Cronbach (HADS, SCOPA-AUT y mPPRS) [102]. En los demás, se utilizó el coeficiente de correlación intraclase derivado del análisis test-retest (CISI-PD, NMSS y MDS-UPDRS).

Para las puntuaciones totales de las escalas, se obtuvo los siguientes valores de EEM: CISI-PD: 2,49; mPPRS: 1,07; SCOPA-AUT: 5,86; HADS-A: 2,64; HADS-D: 2,67. Excepto en el caso de la mPPRS, el EEM resultó menor que ½ DT_{basal}, que es el criterio arbitrario (basado en un índice de fiabilidad mínimo de 0,75) establecido por distintos autores para hablar de una precisión adecuada [84,94]. Sin embargo, para algunos dominios o subescalas de NMSS y SCOPA-AUT, el valor del EEM superó dicho umbral, indicando una relativa falta de precisión.

4.1.5. Valoración global

En la Tabla 4.5 aparece la valoración global de cada escala estudiada en función del resultado de los análisis de sus atributos psicométricos, según los resultados de los estudios propios. Además, y con intención comparativa, se incluye el grado de recomendación que les ha asignado la *Task Force* de la *Movement Disorders Society* en sus revisiones sistemáticas [103].

Tabla 4.5. Evaluación global propia y grado de recomendación de las MDS-Task Forces sobre las escalas validadas en esta Tesis

ESCALA	Calificación global según validación	Grado de recomendación según la MDS-Task Force[a]	Referencia
CISI-PD	Satisfactoria	Tema no revisado	[89]
HADS	Satisfactoria	Ansiedad: Sugerida para *screening* Depresión: Recomendada para *screening*; sugerida para evaluación de gravedad	[90,104,105]
NMSS	Satisfactoria	Recomendada	[94,103,106]
SCOPA-AUT	Aceptable	Recomendada (con limitaciones)	[91,106,107]
mPPRS	Aceptable	Sugerida	[92,108]
MDS-UPDRS	Satisfactoria	Recomendada	[95,109,110]

MDS: Movement Disorder Society.

[a] Una escala es **recomendada** si ha sido utilizada en pacientes con enfermedad de Parkinson, si hay datos publicados en estudios independientemente de los autores originales y si es fiable, válida y sensible al cambio. Una escala es **sugerida** si se ha utilizado en pacientes con enfermedad de Parkinson y si cumple uno de los dos criterios restantes.

4.2. Resultados del estudio de sensibilidad al cambio e interpretabilidad de escalas en enfermedad de Parkinson

Para el estudio de la sensibilidad al cambio e interpretabilidad de medidas clínicas en enfermedad de Parkinson, se han utilizado las escalas CISI-PD, mPPRS, SCOPA-AUT y HADS, que cuentan con datos basales y de seguimiento procedentes del estudio ELEP [93].

Los datos corresponden a 228 pacientes evaluados en 2006-2007 (datos basales) y 2009-2010 (seguimiento), que suponen el 58,91% del total de pacientes incluidos originalmente en el estudio (n=387).

La muestra, de la que algo más de la mitad (120, el 52,6%) eran hombres, tenía una edad media de 62,96 años (desviación típica, DT: 10,61) al comienzo del estudio, y había recibido educación formal durante una media de 10,24 años (DT: 5,56). La distribución basal por estadios de Hoehn y Yahr era la siguiente: 61 pacientes (27,1%) en estadio 1; 116 (51,6%) en estadio 2; 41 (18,2%) en estadio 3 y 7 (3,1%) en estadio 4. La edad media de comienzo de los primeros síntomas reconocibles de la enfermedad fue de 56,19 años (DT: 11,87), y los pacientes tenían una duración media de la enfermedad de 6,78 años (DT: 5,73).

4.1.6. Sensibilidad al cambio

En la Tabla 4.6 se presentan los valores medios basales y de seguimiento de las escalas analizadas en el estudio ELEP. La distribución de pacientes por estadios HY fue significativamente diferente entre las dos evaluaciones. Aunque las escalas validadas, excepto HADS-Ansiedad, mostraron un empeoramiento a lo largo del estudio, este fue estadísticamente significativo solo en el caso de CISI-PD y SCOPA-AUT. Del resto de escalas, solo SCOPA-Motor y la escala visual analógica de dolor mostraron diferencias significativas entre los dos momentos de evaluación.

Tabla 4.6. Puntuaciones medias basales y de seguimiento de las escalas utilizadas en el estudio ELEP.

	Basal (T1)		Seguimiento (T2)		
ESCALA	**Media**	**DT**	**Media**	**DT**	**p[a]**
A. Escalas validadas					
CISI-PD	7,02	3,98	8,79	3,94	**<0,001**
mPPRS	1,06	1,44	1,19	1,52	0,164
SCOPA-Autonómico	19,16	10,89	21,25	10,06	**<0,001**
- Gastrointestinal	5,50	3,94	6,44	3,86	**<0,001**
- Urinario	6,69	4,58	7,83	4,67	**<0,001**
- Cardiovascular	1,27	1,69	1,34	1,71	0,475
- Termorregulatorio	3,27	3,00	3,33	3,01	0,526
- Pupilomotor	1,03	1,15	1,24	1,17	0,012
- Sexual	2,11	1,96	2,45	2,17	0,046
HADS-Ansiedad	7,04	4,02	6,88	4,20	0,644
HADS-Depresión	5,36	3,70	5,85	3,93	0,059
B. Otras escalas del estudio ELEP					
SCOPA-Motor	14,64	7,92	17,88	9,30	**<0,001**
SCOPA-Cognición	25,04	6,14	24,30	7,41	0,055
Dolor	17,95	20,67	22,56	24,27	0,027
Fatiga	24,54	27,15	33,28	29,35	**<0,001**
SCOPA Sueño Nocturno	5,33	4,00	4,56	3,55	0,017
SCOPA Sueño Diurno	3,64	3,09	3,78	3,01	0,453

[a] Test de Wilcoxon. Corrección de Bonferroni: p=0.002

DT: desviación típica.

Estos cambios se correlacionaron con los cambios en los dos aspectos más destacados de la enfermedad: el trastorno motor (representado por la puntuación total de la SCOPA-M) y el trastorno cognitivo (representado por la puntuación total de la SCOPA-COG), que se utilizaron como "anclaje" externo para esta finalidad. El cambio en las puntuaciones de la escala CISI-PD correlacionó 0,61 con el cambio en la escala SCOPA-M. Los coeficientes de correlación del cambio en el resto de escalas analizadas con el cambio en las puntuaciones de SCOPA-M y SCOPA-COG fueron inferiores a 0,30, como se observa en la Tabla 4.7.

Tabla 4.7. Correlaciones del cambio en las escalas analizadas con el cambio en síntomas motores y cognitivos.

ESCALA	SCOPA-M	SCOPA-COG
CISI-PD	0,61**	-0,20**
mPPRS	0,27**	-0,18**
SCOPA-AUT	-0,02	-0,17
HADS-A	0,16*	-0,21
HADS-D	0,22**	-0,06

* $p < 0,05$

** $p < 0,01$

En la Tabla 4.8 aparecen los estadísticos de sensibilidad al cambio: tamaño del efecto y respuesta media estandarizada. Se observa que, del grupo de escalas validadas en el marco del estudio ELEP, la escala CISI-PD fue la que alcanzó un mayor tamaño del efecto (0,44) y respuesta media estandarizada (0,51), mientras que HADS-Ansiedad fue la escala que mostró menores tamaño del efecto y respuesta media estandarizada (0,04 para ambos estadísticos).

Tabla 4.8. Estadísticos de sensibilidad al cambio.

ESCALA	Tamaño del efecto	Respuesta media estandarizada
A. Escalas validadas		
CISI-PD	0,44	0,51
mPPRS	0,09	0,08
SCOPA-Autonómico	0,19	0,27
- Gastrointestinal	0,24	0,28
- Urinario	0,25	0,26
- Cardiovascular	0,04	0,05
- Termorregulatorio	0,02	0,02
- Pupilomotor	0,18	0,18
- Sexual	0,17	0,18
HADS-Ansiedad	0,04	0,04
HADS-Depresión	0,13	0,13
B. Otras escalas del estudio ELEP		
SCOPA-Motor	0,41	0,41
SCOPA-Cognición	0,12	0,15
Dolor	0,22	0,18
Fatiga	0,32	0,28
SCOPA-Sueño Nocturno	0,19	0,19
SCOPA-Sueño Diurno	0,05	0,04

4.1.7. Interpretabilidad

Los estadísticos relativos a la magnitud del cambio aparecen en la Tabla 4.9. En la misma tabla se muestran los resultados para otras variables del estudio.

Tabla 4.9. Coeficientes de magnitud del cambio.

ESCALA	Δ	CR (%)	TAC	TAC (%)	TE	Interpretación TE*
A. Escalas validadas						
CISI-PD	1,79	25,50	0,60	2,49	0,44	Pequeño
mPPRS	0,13	12,26	0,04	0,24	0,09	Insignificante
SCOPA-AUT	3,45	18,01	1,15	1,67	0,19	Insignificante
- Gastrointest	0,94	17,09	0,31	1,49	0,24	Pequeño
- Urinario	1,30	19,43	0,43	2,41	0,25	Pequeño
- Cardiovasc	0,06	4,72	0,02	0,22	0,04	Insignificante
- Termorregul	0,07	2,14	0,02	0,19	0,02	Insignificante
- Pupilomot	0,20	19,42	0,07	2,22	0,18	Insignificante
- Sexual	0,40	18,96	0,13	2,22	0,17	Insignificante
HADS-A	-0,18	-2,56	-0,06	0,29	0,04	Insignificante
HADS-D	0,47	8,77	0,16	0,75	0,13	Insignificante
B. Otras escalas del estudio ELEP						
SCOPA-MOT	3,20	21,86	1,07	1,42	0,41	Pequeño
SCOPA-COG	-0,67	-2,68	-0,22	0,52	0,12	Insignificante
Dolor	4,54	25,29	1,51	1,51	0,22	Pequeño
Fatiga	8,75	35,66	2,92	2,92	0,32	Pequeño
SCOPA SN	-0,75	-14,07	-0,25	1,67	0,19	Insignificante
SCOPA SD	0,08	2,20	0,03	0,15	0,05	Insignificante

Δ: diferencia de medias basal-final; CR: cambio relativo; TAC: tasa anual de cambio; TAC (%): TAC estandarizada sobre el total de la escala; TE: tamaño del efecto.

* Según criterios de Cohen [68].

Las escalas CISI-PD y SCOPA-AUT fueron las que mostraron mayor diferencia entre las puntuaciones basales y de seguimiento según el porcentaje de cambio relativo (25,50% y 18,01%, respectivamente). En el caso de la SCOPA-AUT, las dimensiones urinaria, pupilomotora y sexual mostraron los mayores porcentajes de cambio relativo (19,43%, 19,42% y 18,96%, respectivamente). La tasa anual de cambio estandarizada sobre el total de la escala osciló entre 2,49 (CISI-PD) y 0,24 (mPPRS), con un rango entre 0,19 y 2,41 para las dimensiones de SCOPA-AUT.

En la Tabla 4.10 se muestran los resultados del análisis de interpretabilidad basado en los criterios 1 EEM, ½ DT_{basal} y 10% de la puntuación total.

Tabla 4.10. Valores de interpretabilidad de las escalas.

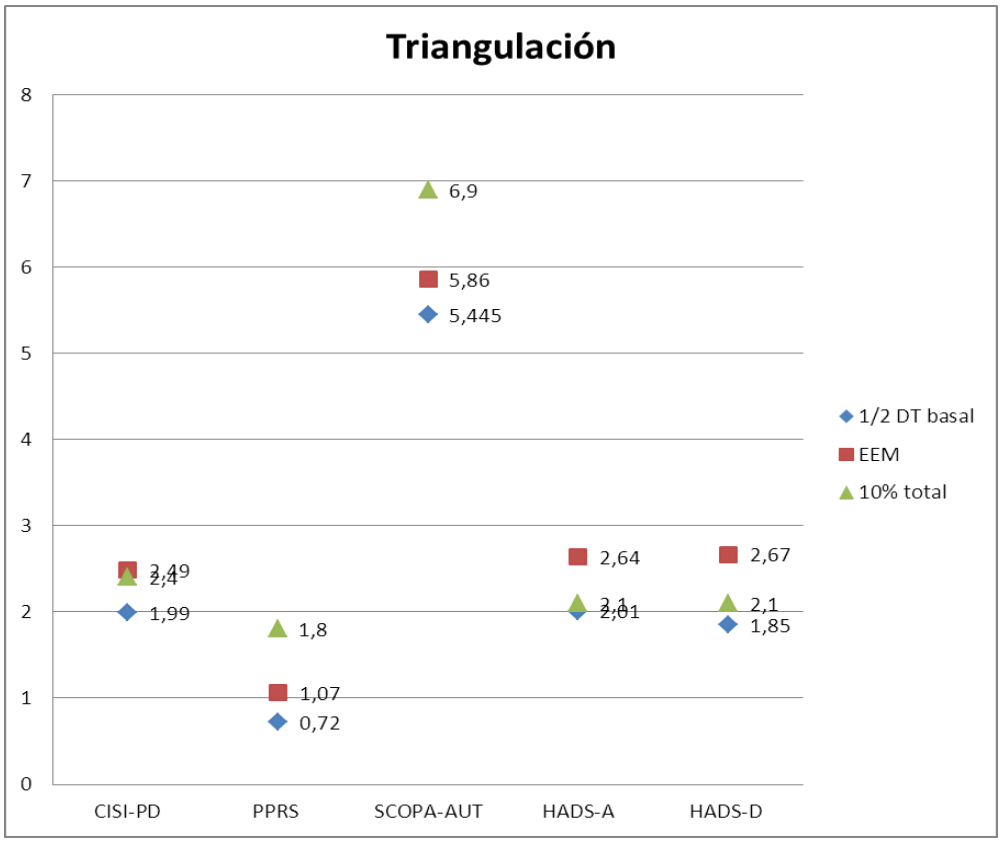
ESCALA	Δ	1 EEM	½ DT_{basal}	10% del total*
CISI-PD	1,79	2,49	1,99	2,4
mPPRS	0,13	1,07	0,72	1,8
SCOPA-Autonómico	3,45	5,86	5,44	6,9
- Gastrointestinal	0,94	2,43	1,97	2,1
- Urinario	1,30	3,11	2,29	1,8
- Cardiovascular	0,06	1,06	0,84	0,9
- Termorregulatorio	0,07	2,06	1,50	1,2
- Pupilomotor	0,20	0,83	0,57	0,3
- Sexual	0,40	1,24	0,98	0,6
HADS-Ansiedad	-0,18	2,64	2,01	2,1
HADS-Depresión	0,47	2,67	1,85	2,1

EEM: error estándar de la medida; DT: desviación típica.

* Valores correspondientes al 10% de la máxima puntuación total teórica de la escala.

Analizando la tabla, se observa que el valor más bajo de las puntuaciones total de cada escala lo ofrece la ½ DT_{basal}. Sin embargo, los demás resultados no siguen ningún otro patrón. De los tres valores para cada escala, se puede calcular un valor medio (al que denominaremos "valor estimado del cambio", VEC) que podría representar el más cercano al "valor real": CISI-PD = 2,29; PPRS = 1,20; SCOPA-AUT = 6,07; HADS-A = 2,25; HADS-D = 2,21. En la Figura 4.1 se muestra la triangulación de los valores de las puntuaciones totales de las escalas.

Figura 4.1. Triangulación de valores indicativos de cambio real.



Basado en Revicki et al. Journal of Clinical Epidemiology 2008;61:102-109

Aplicando el criterio umbral de 1 VEC, se puede identificar a los pacientes que mejoran, empeoran o permanecen estables, tal como se muestra en la Tabla 4.11. Se observa que más de un 50% de la muestra permaneció estable a lo largo del tiempo de estudio en las escalas analizadas. Entre un 13,60% y un 38,79% de los pacientes mostraron un empeoramiento (valores correspondientes a las escalas mPPRS y CISI-

PD, respectivamente). Los valores del número necesario de pacientes a observar para detectar un empeoramiento oscilaron entre 2,62 (CISI-PD) y 7,35 (mPPRS).

Tabla 4.11. Distribución de pacientes en función del cambio* y número necesario de pacientes a observar para detectar un empeoramiento.

ESCALA	Mejoran		Estables		Empeoran		
	N	%	N	%	N	%	NNO†
CISI-PD	20	8,77	120	52,63	87	38,16	2,62
mPPRS	22	9,65	173	75,88	31	13,60	7,35
SCOPA-Autonómico	8	6,90	68	58,62	40	34,50	2,90
HADS-Ansiedad	53	23,25	119	52,19	54	23,68	4,22
HADS-Depresión	43	18,86	127	55,70	56	24,56	4,07

* Se consideró "cambio" a una modificación en la puntuación total basal-final de 1 VEC, tanto en sentido de mejoría como de empeoramiento. Los pacientes con cambios menores de 1 VEC fueron considerados "estables".

† Debido al carácter progresivo de la enfermedad de Parkinson se consideró como cambio de interés el empeoramiento. NNO: número necesario de pacientes a observar para detectar un empeoramiento ≥ 1 VEC (1 / porcentaje de pacientes que empeoraron 1 VEC) x 100.

DISCUSIÓN

5. Discusión

El objetivo de esta Tesis Doctoral es presentar la sistematización de la metodología de validación e interpretación de resultados de medidas clínicas aplicadas a la enfermedad de Parkinson.

La enfermedad de Parkinson es una enfermedad neurodegenerativa compleja que requiere de instrumentos de evaluación que ayuden a identificar la presencia de déficits y complicaciones, documenten la evolución del proceso, cuantifiquen la gravedad e impacto de los síntomas, valoren el efecto de las intervenciones terapéuticas y faciliten el registro de datos y el intercambio de información entre clínicos, con otros profesionales y con los propios pacientes y sus familias. En las últimas décadas se han desarrollado una gran variedad de escalas para la evaluación de las manifestaciones clínicas de la enfermedad de Parkinson y paralelamente, se han publicado importantes directrices para la validación de escalas de resultados informados por los pacientes, o *patient-reported outcomes* (PROs). Estas directrices definen los requisitos que debe cumplir un instrumento de medición del estado de salud para garantizar su calidad: la disponibilidad de información pertinente sobre el modelo conceptual y de medida, la viabilidad y aceptabilidad, los diferentes tipos de validez, la fiabilidad en sus aspectos de consistencia interna y reproducibilidad, además de otros aspectos como la carga administrativa y de respuesta, las formas alternativas y la adaptación lingüística y cultural.

La enfermedad de Parkinson una enfermedad crónica cuyas manifestaciones evolucionan en gravedad y frecuencia a lo largo del curso evolutivo de la enfermedad. Por ello, es de especial interés el desarrollo de instrumentos de evaluación con sensibilidad al cambio y de una metodología que permita la asignación de un significado a dicho cambio, o interpretabilidad.

Esta Tesis presenta los resultados de los estudios de validación (Objetivo 1), y del estudio de sensibilidad al cambio (Objetivo 2) e interpretabilidad (Objetivo 3) de escalas clínicas para la enfermedad de Parkinson

5.1. Estudios de validación de medidas clínicas en enfermedad de Parkinson

En relación al **Objetivo 1**, se han analizado las propiedades psicométricas de cuatro medidas clínicas, las escalas CISI-PD, NMSS, mPPRS y MDS-UPDRS, y de dos medidas de resultados comunicados por los pacientes, las escalas HADS y SCOPA-AUT [89-92,94,95]. Es de destacar que para las escalas NMSS y mPPRS se trata del primer estudio de validación formal realizado en España y para la MDS-UPDRS, el de la versión oficial para el mundo de habla hispana.

Para cada escala, se analizaron las siguientes propiedades: viabilidad y aceptabilidad, fiabilidad (consistencia interna, en todos los casos; reproducibilidad o estabilidad temporal en tres de las escalas), validez de constructo y precisión.

Para la mayor parte de las escalas analizadas, la viabilidad y la aceptabilidad resultaron ser satisfactorias (Tabla 4.2), siguiendo los criterios mostrados en la Tabla 1.4. La escala SCOPA-AUT fue la que presentó un porcentaje de datos computables más bajo debido a la presencia de ítems de índole sexual que con frecuencia generan datos perdidos. La escala mPPRS mostró un importante efecto suelo (46,65%), debido al relativamente bajo porcentaje de pacientes que tenían problemas psiquiátricos en la muestra.

En general, los datos de consistencia interna fueron satisfactorios para las escalas en estudio (Tabla 4.3), si bien algunos parámetros de éste atributo resultaron inferiores al criterio estándar en alguna escala. Es el caso de las escalas NMSS, SCOPA-AUT, mPPRS y MDS-UPDRS, aunque hay que tener en cuenta que se trata de instrumentos que incluyen gran variedad de síntomas agrupados en diversas dimensiones. Teniendo en cuenta esta circunstancia, Clark y colaboradores [61] distinguen entre las escalas de medida de constructos amplios y las que miden constructos más restringidos o delimitados. Las escalas citadas pertenecen al primer grupo, en el que algunos de los criterios de consistencia interna, como el coeficiente de homogeneidad, pueden ser menos estrictos debido a la variedad de constructos incluidos.

La fiabilidad test-retest, analizada en las escalas CISI-PD, NMSS y MDS-UPDRS, superó ampliamente el criterio mínimo de 0,70 en las puntuaciones totales y la mayoría de dominios que componen dichas escalas (Tabla 4.3). Los valores inferiores a 0,70 en el índice kappa del ítem de exploración motora del CISI-PD y el ICC del dominio sexual de la NMSS se han encontrado también en otros estudios [18,25]. Los resultados, por tanto, confirman la satisfactoria reproducibilidad de las escalas analizadas y la estabilidad en sus puntuaciones.

La validez de contenido solo se estudió formalmente en el caso de la mPPRS, con buenos resultados. Para el resto de escalas, la definición explícita de un modelo conceptual y de medida a partir del cual se han desarrollado y la forma de construcción de dichas escalas a partir de revisión bibliográfica y la participación de expertos, pacientes y familiares, avalan su validez de contenido [46].

La validez de constructo, o de comprobación de hipótesis sobre el constructo tal como la ha denominado la iniciativa COSMIN [50], resultó ser satisfactoria en casi todas las escalas (Tabla 4.4). En relación con la validez convergente, los coeficientes de correlación de las escalas con otras medidas utilizadas para evaluar los mismos constructos estuvieron en un rango de entre 0,42 y 0,85, confirmándose así la potencial relación de la escala en estudio con el constructo que pretende medir. Es particularmente interesante la estrecha relación entre las escalas analizadas y las medidas de calidad de vida genéricas (EQ-5D) y específicas (PDQ-39, PDQ-8, SCOPA-PS) acompañantes (Tabla 4.4), ya que los constructos evaluados por estas escalas se han reconocido como determinantes de la calidad de vida de los pacientes [10,11].

Como consecuencia de la progresión en la patología subyacente a la enfermedad de Parkinson, acontece un empeoramiento paulatino del estado de salud del paciente. Tal deterioro se debe reflejar en las puntuaciones de las medidas clínicas y de resultados comunicados por los pacientes en los distintos estadios de la enfermedad [6,111]. Esta hipótesis se confirmó mediante el análisis de la validez para grupos conocidos: todas las escalas distinguieron entre pacientes agrupados según la clasificación en estadios de Hoehn y Yahr, con diferencias en las puntuaciones estadísticamente significativas. En algunas escalas, como la mPPRS, SCOPA-AUT y MDS-UPDRS, se detectaron diferencias también en función de la edad y duración de la enfermedad, pero ese hallazgo no es unívoco. Aunque, de manera global, la enfermedad de Parkinson

empeora con la edad y la duración de la enfermedad, la relación entre su curso evolutivo y estos factores no es muy estrecha como consecuencia de la diferente edad de comienzo, variabilidad entre individuos, subtipo de enfermedad y otros factores [112–115].

Respecto a la validez estructural o interna (Tabla 4.4), los análisis factoriales realizados apoyan la agrupación de los ítems en las subescalas o dominios de los instrumentos CISI-PD, HADS y mPPRS. Hay que tener en cuenta, sin embargo, que sólo en el caso de CISI-PD y mPPRS se comprobó la unidimensionalidad de las mismas mediante análisis factorial confirmatorio. La validez interna de las escalas NMSS, SCOPA-AUT y MDS-UPDRS se puso a prueba mediante análisis de las correlaciones inter-dominios, en el que se detectaron coeficientes inferiores a 0,30. Estas bajas correlaciones tienen que ver con la diversidad de dominios incluidos, algunos de los cuales tienen escasa relación entre sí. No obstante, se diseñaron y utilizan combinados en una misma escala por el interés de disponer de medidas que permitan evaluar diversos aspectos de manera práctica y simultáneamente. Las múltiples manifestaciones que pueden estar presentes en los pacientes con enfermedad de Parkinson ha propiciado el desarrollo y uso de instrumentos específicos para un síntoma o trastorno (por ejemplo, la *Parkinson Anxiety Scale*, PAS) [116] y de instrumentos "comprehensivos", para identificar la presencia o evaluar un elevado número de trastornos (por ejemplo, la MDS-UPDRS, el Cuestionario de Síntomas no Motores, NMSQuest, y la NMSS) [24,25].

Un aspecto muy relacionado con la fiabilidad de las escalas estudiadas es la precisión. En todas las escalas analizadas, excepto en la mPPRS, la precisión resultó ser satisfactoria, con un EEM menor que el criterio de ½ DT_{basal} [117]. Esto indica que las escalas estudiadas tienen la capacidad de distinguir pequeñas diferencias dentro de su rango de medida [85]. El EEM marca el nivel a partir del cual un cambio en la puntuación de una medida puede considerarse "real" (mínimo cambio detectable) [118], es decir, supera el "ruido" atribuible al mero error de medida [85]. En este sentido, puede considerarse como un vínculo entre la fiabilidad de una escala y su sensibilidad al cambio (Objetivo 2 de la Tesis) [71], con la ventaja de que es relativamente estable entre poblaciones, ya que es una característica del instrumento y no de la muestra [69]. El EEM es también una de las técnicas de cálculo aplicable a la interpretabilidad

(Objetivo 3) basada en la distribución de puntuaciones para calcular el mínimo cambio importante [34].

En general, las propiedades psicométricas de las escalas analizadas resultaron satisfactorias y el juicio sobre su calidad resultó, en conjunto, coincidente con el grado de recomendación de la MDS-Task Force (Tabla 4.5). Los resultados de los estudios de validación presentados en esta Tesis han sido publicados en diversas revistas científicas internacionales [89-92,94,95].

5.2. Estudio de sensibilidad al cambio e interpretabilidad de las medidas clínicas en enfermedad de Parkinson

Para el **Objetivo 2**, se calculó la sensibilidad al cambio de las escalas para las que se contaba con datos de seguimiento a 4 años mediante el contraste estadístico de las diferencias entre las puntuaciones de la evaluación basal y la de seguimiento, la correlación del cambio en las puntuaciones con el cambio en otras medidas y los estadísticos cambio relativo y respuesta media estandarizada.

Aunque el cambio detectado fue en la dirección prevista (empeoramiento) teniendo en cuenta que se trata de una enfermedad progresiva, solo se observó un empeoramiento estadísticamente significativo en las puntuaciones de las escalas CISI-PD y SCOPA-AUT (puntuación total y dominios gastrointestinal y urinario) (Tabla 4.6). Esto puede deberse, una vez establecida la adecuada sensibilidad de los instrumentos en estudio, a que el período de observación no fue lo suficientemente largo como para detectar cambios significativos en todas las escalas, a que la duración de la permanencia en un estadio difiere de unos estadios a otros (por lo cual, la composición de la muestra influye estos resultados) y a que las distintas manifestaciones de la enfermedad de Parkinson evolucionan a un ritmo diferente [119-121]. En la evaluación basal del estudio longitudinal, el 70% de la muestra se encontraba en estadios HY 2 y 3, que son de larga permanencia, por lo cual, las expectativas de cambio son menores que en otros estadios [122]. Hay evidencia de que el temblor, los déficits cognitivos y la depresión empeoran más lentamente que otras manifestaciones [123,124], por lo que el tiempo de observación para detectar cambios en algunos trastornos debe ser más prolongado. Las bajas correlaciones observadas entre el cambio en las puntuaciones de las escalas analizadas y las escalas SCOPA-M y SCOPA-COG (excepto las de CISI-PD

con SCOPA-M) se podrían explicar por estas diferencias en la progresión de los síntomas (Tabla 4.7). Por otra parte, los efectos beneficiosos de algunos tratamientos podrían compensar el incremento en intensidad de algunas manifestaciones durante tiempo prolongado, contribuyendo de este modo a la neutralización del efecto de la progresión.

Los estadísticos relativos a la sensibilidad al cambio (Tabla 4.8) mostraron, de forma congruente con lo observado hasta ahora, valores bajos en general, indicando que durante el período de observación solo hubo pequeños cambios o, en su defecto, una baja sensibilidad al cambio de las escalas analizadas. Sin embargo, algunas de las escalas como la NMSS han sido utilizadas en estudios con intervenciones y han mostrado una sensibilidad al cambio adecuada [125–127].

Los estadísticos tamaño del efecto y cambio relativo forman parte de las técnicas basadas en la distribución de las puntuaciones. El tamaño del efecto es una medida de sensibilidad al cambio ampliamente usada en ensayos clínicos ya que cuenta con unos valores estándar que facilitan su interpretación [68,128] y, además, aporta información sobre la magnitud del cambio que resulta útil para la interpretación de resultados. La respuesta media estandarizada proporciona información del cambio en relación con la desviación típica del mismo [128], y se interpreta con los mismos criterios que el tamaño del efecto [69] aunque ambas técnicas arrojan resultados diferentes. La mayor ventaja de estos estadísticos, aparte la estandarización, es que incluyen en su fórmula la variabilidad (de la muestra o del cambio) mediante desviaciones típicas. Su mayor limitación es que, al no estar referidos a un criterio externo, no es posible determinar si el cambio que reflejan se debe a un cambio real en el constructo que se está evaluando [66]. En este sentido, Husted y colaboradores distinguen la sensibilidad al cambio "interna" (la capacidad de detectar un cambio en el tiempo) de la "externa" (relación entre los cambios de una medida con los cambios en otra medida de referencia) [128]. El tamaño del efecto y la respuesta media estandarizada pertenecen al primer grupo (sensibilidad al cambio interna).

La interpretación del cambio se puede establecer con las técnicas utilizadas para responder al **Objetivo 3** de esta Tesis: determinación de la magnitud del cambio (Δ, CR, TAC), de los umbrales de cambio mediante métodos basados en la distribución de las puntuaciones de las escalas (1 EEM, ½ $DT_{basal,}$ 10% del máximo teórico) y de la

importancia del cambio (porcentaje de pacientes que mejoran, permanecen estables o empeoran y número necesario de pacientes a observar para detectar un empeoramiento). Aunque la mayor parte de los métodos para establecer la sensibilidad al cambio son también utilizables como técnicas para la interpretación de los resultados (y para validación longitudinal) [66] nuestra propuesta se basa en la utilización de una sistemática sencilla y comprensible que permita contestar a la pregunta: ¿es importante el cambio observado?. Con demasiada frecuencia, los resultados de intervenciones terapéuticas, ensayos clínicos y estudios de seguimiento se limitan a unos datos estadísticos, sin que se informe sobre el impacto que dicho cambio ha ejercido sobre el estado de salud y calidad de vida de los pacientes. Por otra parte, no existe un método único que sea ampliamente reconocido para determinar la importancia del cambio. Nuestro modelo de interpretación de resultados para grupos, basado en análisis estadísticos de distribución, se basa en dos principios lógicos:

1. El cambio será tanto más importante cuanto mayor sea su magnitud.
2. Un cambio será más importante cuando afecte a mayor proporción de pacientes.

En la aplicación llevada a cabo para esta Tesis, el cambio fue producido por la progresión de la enfermedad, que acontece a pesar de los tratamientos paliativos, aunque estos tienden a enmascarar la evolución natural del trastorno. Los valores de magnitud del cambio indican que CISI-PD y SCOPA-AUT mostraron las mayores diferencias entre la evaluación basal y el seguimiento, si bien los cambios en las escalas son de pequeña dimensión en general, como se aprecia en la Tabla 4.9. Las puntuaciones cambiaron entre un -2,56% (HADS-A) y un 25,50% (CISI-PD), como refleja el porcentaje de cambio relativo, mientras que la tasa anual de cambio estandarizada nos indicó un empeoramiento anual en las puntuaciones de entre 0,17% (mPPRS) y 1,87% (CISI-PD). El principal problema de estos parámetros es la falta de puntos de corte que sirvan de criterio para definir el cambio como pequeño, moderado o grande, como sí ocurre con el tamaño del efecto. Por ello, solo adquieren significado cuando se comparan con el cambio de otras escalas o con un valor esperado. Además, la tasa anual de cambio asume que dicho cambio es lineal, por lo que no refleja la progresión natural de la mayoría de enfermedades [69].

En el caso del tamaño del efecto, los puntos de corte señalados por Cohen [68] se han adoptado como criterio de consenso para interpretar la magnitud del cambio. En el caso de las escalas estudiadas, el tamaño del efecto fue insignificante o de pequeña magnitud (Tabla 4.9). Esta medida se expresa en unidades estandarizadas, por lo que permite comparar resultados de diversas escalas, y se ha propuesto por diversos autores como la medida de sensibilidad al cambio más adecuada [129], y por extensión, de interpretación del cambio. Como se ha comentado anteriormente, se desconoce hasta qué punto refleja un cambio real al no estar referido a un criterio externo o anclaje, además de estar muy influido por la distribución de las puntuaciones en la evaluación basal [130]. No obstante, las escalas que presentaban un porcentaje de cambio relativo más alto, CISI-PD y SCOPA-AUT, son las que presentaron también valores más altos de tamaño del efecto, hallazgo indicativo de un acuerdo entre ambos métodos.

Para estimar el umbral de un mínimo cambio importante se han utilizado tres parámetros: el error estándar de medida (EEM), ½ DT_{basal} y el 10% de la puntuación total de la escala [79–85] (Tabla 4.10). El EEM es una medida de precisión de la escala que también puede utilizarse como estimación del mínimo cambio importante, basándose en investigaciones que apoyan el hecho de que 1 EEM equivale a una diferencia aproximada de 0,5 en una escala de siete puntos y a ½ DT_{basal} cuando el coeficiente de fiabilidad es de 0,75 o más [34,85] Por otro lado, la mayor parte de los valores de mínimo cambio importante publicados en distintos estudios son equivalentes al 10% de la puntuación total de la escala [82]. Sin embargo, las tres medidas de interpretación del cambio que se muestran en la Tabla 4.10 arrojan valores ligeramente diferentes, sin que exista un patrón o regla que indique que una sea más adecuada que otra. Por ello se ha recurrido al análisis de la triangulación de los valores, presentado en la Figura 4.1. Para CISI-PD, por ejemplo, el cambio real oscilaría entre 1,99 (½ DT_{basal}) y 2,49 (1 EEM), y podría ser más próximo al promedio de los tres parámetros (valor estimado del cambio = 2,29). Hasta qué punto este cambio es significativo para el paciente es algo que, sin embargo, no es directamente deducible de las técnicas basadas en la distribución de puntuaciones puesto que no se cuenta con la información proveniente del mismo paciente registrada por métodos basados en anclaje [118,131]. Sin embargo, el valor estimado del cambio ofrece una aproximación al mínimo cambio importante que puede tener ventajas, ya que los métodos de anclaje tienen también considerables inconvenientes: las escalas de transición están compuestas por un solo

ítem, a pesar de que el constructo explorado sea complejo; la elección del anclaje puede no ser el adecuado; los niveles de cambio mínimo, moderado y grande son arbitrarios, dependen de la subjetividad del paciente y carecen de control clínico; y por último, se basan en una evaluación retrospectiva que tiene un componente de error [132].

Cualquiera de estos valores (por ejemplo, 1 EEM o ½ DT_{basal}), u otros prefijados, podría usarse para clasificar a los pacientes que mejoran, empeoran o permanecen estables a lo largo del tiempo, y del número necesario de pacientes que hay que observar para detectar un empeoramiento, como se muestra en la Tabla 4.11. Se trata de un sistema muy intuitivo y sencillo para interpretar los resultados a simple vista. En el presente estudio, utilizando el criterio de 1 VEC (como mínimo cambio real), se observa que entre un 14 y un 38% de los pacientes empeoró en los constructos estudiados, sobre todo en las puntuaciones de CISI-PD y SCOPA-AUT, lo que es congruente con la mayor sensibilidad al cambio en estas escalas y con la evolución natural de la enfermedad de Parkinson. De hecho, solo son necesarios 2,6 pacientes para observar un empeoramiento con estas escalas, en comparación con los 7,35 que se necesitan para la mPPRS.

5.3. Limitaciones

Es necesario señalar una serie de limitaciones en los estudios referidos en esta Tesis. En primer lugar, en todos los estudios se utilizó una muestra de conveniencia, no poblacional ni aleatoria, compuesta por pacientes ambulatorios atendidos en distintas unidades o centros especializados en la atención a pacientes con enfermedad de Parkinson o trastornos del movimiento, tanto en nuestro país como en otros. Precisamente por ello, su composición es característica de la población con enfermedad de Parkinson atendida en el entorno clínico especializado: predominio de pacientes en estadios leve/moderado de la clasificación Hoehn y Yahr, y muy pocos en los estadios avanzados.

En los estudios de validación no se incluyeron algunos de los atributos de las escalas como son la fiabilidad test-retest (ausente en el caso de las escalas HADS, SCOPA-AUT y mPPRS) y la validez de criterio. Es necesario tener en cuenta que para

la validez de criterio se necesita una medida de referencia o *gold standard* que no existe para las escalas que se han analizado. El estudio formal de la validez de contenido de las escalas solo se ha llevado a cabo en el caso de la mPPRS, aunque para el resto de escalas, dicha validez se apoya en el proceso de construcción de las mismas.

Por último, como ya se ha señalado en la Discusión en relación con las medidas de sensibilidad al cambio e interpretación de los resultados, el período de seguimiento de los pacientes puede no haber sido lo suficientemente largo como para detectar cambios significativos en las escalas. Sin embargo, existen distintos estudios que avalan la sensibilidad al cambio como resultado de una intervención de algunas de las escalas, como es el caso de la NMSS.

5.4. Implicaciones

El desarrollo y validación de escalas clínicas y de resultados comunicados por los pacientes es imprescindible para la práctica clínica y para la investigación en todos los ámbitos clínicos. En pacientes con enfermedad de Parkinson, la complejidad y el carácter subjetivo de muchas de sus manifestaciones hace aún más necesario el disponer de medidas validadas que permitan acceder a una información y estimaciones que de otra manera no sería posible obtener. Sin embargo, en los últimos años hemos asistido a una proliferación de guías y directrices sobre validación de escalas, con distintas terminologías y enfoques, lo que puede llevar a confusión. La sistematización de una metodología de validación de escalas presentada en esta Tesis es clave para disponer de un listado de propiedades métricas y criterios estándar sobre los que existe consenso y guiar así la elaboración y validación de escalas.

Además de ser fiables, válidas y precisas, las escalas clínicas y de resultados comunicados por los pacientes deben ser sensibles a los cambios en el estado clínico de los pacientes, proporcionando una información fácilmente interpretable y significativa para profesionales sanitarios y pacientes, sobre todo cuando estas se utilizan como referencia a la hora de tomar decisiones clínicas.

CONCLUSIONES

6. Conclusiones

Este estudio permite obtener las siguientes conclusiones en función de los tres Objetivos propuestos:

6.1. Conclusiones del Objetivo 1

1. La validación de escalas clínicas y de resultados comunicados por los pacientes en enfermedad de Parkinson puede seguir los mismos principios, métodos y estándares que guían la validación de medidas en otros ámbitos.

2. Existe consenso en la necesidad de calcular e informar de, al menos, las siguientes propiedades psicométricas: viabilidad y aceptabilidad; fiabilidad, entendida como consistencia interna y reproducibilidad; validez y sus tipos, validez de contenido y validez de constructo; precisión y sensibilidad al cambio.

3. Las escalas analizadas en esta Tesis han mostrado unos valores de viabilidad y aceptabilidad entre aceptables y satisfactorios, lo que indica que en general recogen todo el espectro de intensidad del constructo evaluado y que son aplicables en pacientes con enfermedad de Parkinson.

4. Los resultados satisfactorios del análisis de la consistencia interna, muestran que los ítems de las escalas validadas en esta Tesis miden el mismo constructo. En el caso de las escalas globales multi-dominio como NMSS, SCOPA-AUT, mPPRS y MDS-UPDRS, los criterios de consistencia interna deben ser menos rigurosos.

5. La fiabilidad test-retest o estabilidad temporal se analizó en las escalas CISI-PD, NMSS y MDS-UPDRS con resultados adecuados, lo que indica que las puntuaciones de dichas escalas son estables, atributo que contribuye a que la precisión sea adecuada.

6. La validez de contenido, si bien no se analizó formalmente excepto en la mPPRS, queda garantizada por el proceso de construcción de cada una de las escalas.

7. Las hipótesis relativas a la validez de constructo se han visto confirmadas en su mayor parte, particularmente las relativas a la validez convergente y la validez para grupos conocidos. Las distintas técnicas utilizadas para comprobar la validez estructural o interna de las escalas analizadas avalan, en general, la composición y estructura de las mismas.

8. Por último, se demostró una precisión satisfactoria para la mayor parte de las escalas, excepto la mPPRS, lo que indica que son capaces de distinguir pequeñas diferencias y que sus puntuaciones presentan un error de medida relativamente bajo.

9. En resumen, las escalas clínicas CISI-PD, NMSS y MDS-UPDRS y las medidas de resultados comunicados por los pacientes SCOPA-AUT y HADS han mostrado propiedades psicométricas satisfactorias, en consonancia con el grado de recomendación asignado por las respectivas revisiones de la *MDS Task Force*. En solo una de las escalas analizadas, la mPPRS, nuestra valoración global ha sido "aceptable" en lugar de "satisfactoria" y coincide con una recomendación de nivel moderado por parte de la *MDS Task Force*.

6.2. Conclusiones del Objetivo 2

1. Las técnicas de sensibilidad al cambio basadas en la distribución de las puntuaciones utilizadas en esta Tesis (tamaño del efecto y respuesta media estandarizada) permiten comprobar la capacidad de un instrumento de distinguir cambios en el constructo que se está evaluando (validez longitudinal).

2. Al analizar las diferencias en las puntuaciones de las escalas CISI-PD, mPPRS, SCOPA-AUT y HADS en dos momentos temporales se observó una tendencia al empeoramiento de las mismas, congruente con la evolución

natural de la enfermedad de Parkinson. Dichas diferencias solo fueron significativas en el caso de las escalas CISI-PD y SCOPA-AUT.

3. Los estadísticos de sensibilidad al cambio utilizados indicaron que los cambios fueron pequeños, debido a una muestra en estadios intermedios de la enfermedad, un período de observación relativamente corto o a que las escalas analizadas mostraron baja sensibilidad al cambio.

6.3. Conclusiones del Objetivo 3

1. La importancia del cambio en las puntuaciones de las escalas analizadas se puede valorar mediante métodos que estimen su magnitud (diferencias en las puntuaciones, cambio relativo, tasa anual de cambio, etc.), umbrales de cambio basados en la distribución de las puntuaciones (1 EEM, ½ DT_{basal}, 10% del máximo teórico) y otros cálculos (porcentaje de pacientes que mejoran, permanecen estables o empeoran y número necesario de pacientes a observar para detectar un empeoramiento).

2. Las escalas CISI-PD y SCOPA-AUT fueron las que mostraron las mayores diferencias en las puntuaciones entre la evaluación basal y el seguimiento, aunque los valores indican que el cambio fue relativamente pequeño en general.

3. Los diferentes métodos de cálculo de los umbrales de cambio en las escalas analizadas arrojan valores diferentes. En ausencia de evidencia que indique la mayor conveniencia de una técnica sobre otra, proponemos la triangulación por promedio de las mismas.

4. Los umbrales de cambio permiten clasificar a los pacientes en función de su mejoría, empeoramiento o estabilidad o calcular el número de pacientes a observar para detectar un evento (mejoría o empeoramiento), lo que facilita la interpretación de los resultados longitudinales.

BIBLIOGRAFÍA

7. Bibliografía

1. Gibb WR, Lees AJ. The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. J Neurol Neurosurg Psychiatry. 1988;51(6):745–52.

2. Benito-León J, Bermejo-Pareja F, Rodríguez J, Molina J-A, Gabriel R, Morales J-M, et al. Prevalence of PD and other types of parkinsonism in three elderly populations of central Spain. Mov Disord. 2003;18(3):267–74.

3. Bergareche A, De La Puente E, López de Munain A, Sarasqueta C, de Arce A, Poza JJ, et al. Prevalence of Parkinson's disease and other types of Parkinsonism. A door-to-door survey in Bidasoa, Spain. J Neurol. 2004;251(3):340–5.

4. Del Barrio JL, de Pedro-Cuesta J, Boix R, Acosta J, Bergareche A, Bermejo-Pareja F, et al. Dementia, stroke and Parkinson's disease in Spanish populations: a review of door-to-door prevalence surveys. Neuroepidemiology. 2005;24(4):179–88.

5. García-Ramos R, López Valdés E, Ballesteros L, Jesús S, Mir P. Informe de la Fundación del Cerebro sobre el impacto social de la enfermedad de Parkinson en España. Neurol Barc Spain. 2013;

6. Stern MB, Lang A, Poewe W. Toward a redefinition of Parkinson's disease. Mov Disord. 2012;27(1):54–60.

7. Aarsland D, Brønnick K, Alves G, Tysnes OB, Pedersen KF, Ehrt U, et al. The spectrum of neuropsychiatric symptoms in patients with early untreated Parkinson's disease. J Neurol Neurosurg Psychiatry. 2009;80(8):928–30.

8. Noyce AJ, Bestwick JP, Silveira-Moriyama L, Hawkes CH, Knowles CH, Hardy J, et al. PREDICT-PD: identifying risk of Parkinson's disease in the community: methods and baseline results. J Neurol Neurosurg Psychiatry. 2014;85(1):31–7.

9. Martinez-Martin P, Schapira AHV, Stocchi F, Sethi K, Odin P, MacPhee G, et al. Prevalence of nonmotor symptoms in Parkinson's disease in an international setting; study using nonmotor symptoms questionnaire in 545 patients. Mov Disord. 2007;22(11):1623–9.

10. Barone P, Antonini A, Colosimo C, Marconi R, Morgante L, Avarello TP, et al. The PRIAMO study: A multicenter assessment of nonmotor symptoms and their impact on quality of life in Parkinson's disease. Mov Disord. 2009;24(11):1641–9.

11. Martinez-Martin P, Rodriguez-Blazquez C, Kurtis MM, Chaudhuri KR. The impact of non-motor symptoms on health-related quality of life of patients with Parkinson's disease. Mov Disord. 2011;26(3):399–406.

12. Chaudhuri KR, Odin P, Antonini A, Martinez-Martin P. Parkinson's disease: the non-motor issues. Parkinsonism Relat Disord. 2011;17(10):717–23.

13. Antonini A, Barone P, Marconi R, Morgante L, Zappulla S, Pontieri FE, et al. The progression of non-motor symptoms in Parkinson's disease and their contribution to motor disability and quality of life. J Neurol. 2012;

14. Xu J, Gong DD, Man CF, Fan Y. Parkinson's disease and risk of mortality: meta-analysis and systematic review. Acta Neurol Scand. 2014;129(2):71–9.

15. Fahn S, Elton R, UPDRS program members. Unified Parkinson's disease rating scale. In: Fahn S, Marsden C, Goldstein M, Calne D, editors. Recent developments in Parkinson's disease. Florham Park, NJ: Macmillan Healthcare Information; 1987. p. 153–63.

16. Goetz CG, Fahn S, Martinez-Martin P, Poewe W, Sampaio C, Stebbins GT, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Process, format, and clinimetric testing plan. Mov Disord. 2007;22(1):41–7.

17. Hoehn MM, Yahr MD. Parkinsonism: onset, progression and mortality. Neurology. 1967;17(5):427–42.

18. Marinus J, Visser M, Stiggelbout AM, Rabey JM, Martínez-Martín P, Bonuccelli U, et al. A short scale for the assessment of motor impairments and disabilities in Parkinson's disease: the SPES/SCOPA. J Neurol Neurosurg Psychiatry. 2004;75(3):388–95.

19. Schwab R, England A. Third symposium for Parkinson's disease. Edingburg: Livingstone; 1969. 152-157 p.

20. Stacy M, Hauser R. Development of a Patient Questionnaire to facilitate recognition of motor and non-motor wearing-off in Parkinson's disease. J Neural Transm. 2007;114(2):211–7.

21. Stacy MA, Murphy JM, Greeley DR, Stewart RM, Murck H, Meng X. The sensitivity and specificity of the 9-item Wearing-off Questionnaire. Parkinsonism Relat Disord. 2008;14(3):205–12.

22. Martinez-Martin P, Hernandez B. The Q10 questionnaire for detection of wearing-off phenomena in Parkinson's disease. Parkinsonism Relat Disord. 2012;18(4):382–5.

23. Goetz CG, Nutt JG, Stebbins GT. The Unified Dyskinesia Rating Scale: presentation and clinimetric profile. Mov Disord. 2008;23(16):2398–403.

24. Chaudhuri KR, Martinez-Martin P, Schapira AHV, Stocchi F, Sethi K, Odin P, et al. International multicenter pilot study of the first comprehensive self-completed nonmotor symptoms questionnaire for Parkinson's disease: the NMSQuest study. Mov Disord. 2006;21(7):916–23.

25. Chaudhuri KR, Martinez-Martin P, Brown RG, Sethi K, Stocchi F, Odin P, et al. The metric properties of a novel non-motor symptoms scale for Parkinson's disease: Results from an international pilot study. Mov Disord. 2007;22(13):1901–11.

26. Marinus J, Visser M, Verwey NA, Verhey FRJ, Middelkoop HAM, Stiggelbout AM, et al. Assessment of cognition in Parkinson's disease. Neurology. 2003;61(9):1222–8.

27. Marinus J, Visser M, van Hilten JJ, Lammers GJ, Stiggelbout AM. Assessment of sleep and sleepiness in Parkinson disease. Sleep. 2003;26(8):1049–54.

28. Visser M, Marinus J, Stiggelbout AM, van Hilten JJ. Assessment of autonomic dysfunction in Parkinson's disease: the SCOPA-AUT. Mov Disord. 2004;19(11):1306–12.

29. Visser M, Verbaan D, van Rooden SM, Stiggelbout AM, Marinus J, van Hilten JJ. Assessment of psychiatric complications in Parkinson's disease: The SCOPA-PC. Mov Disord. 2007;22(15):2221–8.

30. Brown RG, Dittner A, Findley L, Wessely SC. The Parkinson fatigue scale. Parkinsonism Relat Disord. 2005;11(1):49–55.

31. U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. Health Qual Life Outcomes. 2006;4:79.

32. Acquadro C, Berzon R, Dubois D, Leidy NK, Marquis P, Revicki D, et al. Incorporating the patient's perspective into drug development and communication: an ad hoc task force report of the Patient-Reported Outcomes (PRO) Harmonization Group meeting at the Food and Drug Administration, February 16, 2001. Value Health. 2003;6(5):522–31.

33. Bowling A. Measuring health. A review of quality of life measurement scales. 3rd edition. Berkshire, UK: Open University Press; 2005. 1-5 p.

34. Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. J Clin Epidemiol. 2008;61(2):102–9.

35. Chassany O, Sagnier P, Marquis P, Fullerton S, Aaronson N. Patient-Reported Outcomes: The Example of Health-Related Quality of Life—A European Guidance Document for the Improved Integration of Health-Related Quality of Life Assessment in the Drug Regulatory Process. Drug Inf J. 2002;36(1):209–38.

36. Committee for Medicinal Products for Human Use. Reflection paper on the regulatory guidance for the use of health-related quality of life (HRQL) measures in the evaluation of medicinal products. European Medicines Agency; 2005. Report No.: EMEA/CHMP/EWP/139391/2004.

37. Calvert M, Brundage M, Jacobsen PB, Schünemann HJ, Efficace F. The CONSORT Patient-Reported Outcome (PRO) extension: implications for clinical trials and practice. Health Qual Life Outcomes. 2013;11(1):184.

38. Gnanasakthy A, Mordin M, Clark M, DeMuro C, Fehnel S, Copley-Merriman C. A review of patient-reported outcome labels in the United States: 2006 to 2010. Value Health. 2012;15(3):437–42.

39. Gnanasakthy A, DeMuro C, Clark M, Mordin M, Thomas S. Role of Patient-Reported Outcome Measures in the Assessment of Central Nervous System Agents. Ther Innov Regul Sci. 2013;47(5):613–8.

40. Gnanasakthy A, Lewis S, Clark M, Mordin M, DeMuro C. Potential of patient-reported outcomes as nonprimary endpoints in clinical trials. Health Qual Life Outcomes. 2013;11:83.

41. Doward LC, Gnanasakthy A, Baker MG. Patient reported outcomes: looking beyond the label claim. Health Qual Life Outcomes. 2010;8:89.

42. Martínez-Martín P, Benito-León J, Alonso F, Catalán MJ, Pondal M, Zamarbide I. Health-related quality of life evaluation by proxy in Parkinson's disease: approach using PDQ-8 and EuroQoL-5D. Mov Disord Off J Mov Disord Soc. 2004 Mar;19(3):312–8.

43. Stevens SS. On the Theory of Scales of Measurement. Science. 1946;103(2684):677–80.

44. DeVellis RF. Classical test theory. Med Care. 2006;44(11 Suppl 3):S50–9.

45. Martinez-Martin P, Rodriguez-Blazquez C, Frades-Payo B. Specific patient-reported outcome measures for Parkinson's disease: analysis and applications. Expert Rev Pharmacoecon Outcomes Res. 2008;8(4):401–18.

46. Terwee CB, Bot SDM, de Boer MR, van der Windt DAWM, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007;60(1):34–42.

47. Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Med Care. 2000;38(9 Suppl):II28–42.

48. Andrich D. Rating scales and Rasch measurement. Expert Rev Pharmacoecon Outcomes Res. 2011;11(5):571–85.

49. Nunnally JC, Bernstein IH. Psychometric theory. New York: McGraw Hill; 1994.

50. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. Qual Life Res. 2010;19(4):539–49.

51. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: attributes and review criteria. Qual Life Res. 2002;11(3):193–205.

52. Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol DL, et al. The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. J Clin Epidemiol. 2010;63(7):737–45.

53. Valderas JM, Ferrer M, Mendívil J, Garin O, Rajmil L, Herdman M, et al. Development of EMPRO: a tool for the standardized assessment of patient-reported outcome measures. Value Health. 2008;11(4):700–8.

54. Fitzpatrick R, Davey C, Buxton MJ, Jones DR. Evaluating patient-based outcome measures for use in clinical trials. Health Technol Assess. 1998;2(14):i – iv, 1–74.

55. Smith SC, Lamping DL, Banerjee S, Harwood R, Foley B, Smith P, et al. Measurement of health-related quality of life for people with dementia: development of a new instrument (DEMQOL) and an evaluation of current methodology. Health Technol Assess. 2005;9(10):1–93, iii – iv.

56. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res. 1995;4(4):293–307.

57. Hays RD, Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life measures. Qual Life Res. 1993;2(6):441–9.

58. Van der Linden FAH, Kragt JJ, Klein M, van der Ploeg HM, Polman CH, Uitdehaag BMJ. Psychometric evaluation of the multiple sclerosis impact scale (MSIS-29) for proxy use. J Neurol Neurosurg Psychiatry. 2005;76(12):1677–81.

59. Ware JE Jr, Gandek B. Methods for testing data quality, scaling assumptions, and reliability: the IQOLA Project approach. International Quality of Life Assessment. J Clin Epidemiol. 1998;51(11):945–52.

60. Eisen M, Ware JE, Donald CA, Brook RH. Measuring components of children's health status. Med Care. 1979;17(9):902–21.

61. Clark LA, Watson D. Constructing validity: Basic issues in objective scale development. Psychol Assess. 1995;7(3):309–19.

62. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.

63. Lynn MR. Determination and quantification of content validity. Nurs Res. 1986;35(6):382–5.

64. Hobart J, Lamping D, Fitzpatrick R, Riazi A, Thompson A. The Multiple Sclerosis Impact Scale (MSIS-29): a new patient-based outcome measure. Brain. 2001;124(Pt 5):962–73.

65. Fayers P, Machin D. Quality of Life. Chichester: Wiley; 2000. 51 p.

66. Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PMM. On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. Qual Life Res. 2003;12(4):349–62.

67. Altman DG. Practical statistics for medical research. Boca Raton, Fl: Chapman & Hall; 1991.

68. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. New York: Academic Press; 1988.

69. Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. J Clin Epidemiol. 2003;56(5):395–407.

70. Armitage P, Berry G, Matthews J n. s. Clinical Trials. Statistical Methods in Medical Research [Internet]. Blackwell Science Ltd; 2002 [cited 2014 Apr 16]. p. 591–647. Available from: http://onlinelibrary.wiley.com/doi/10.1002/9780470773666.ch18/summary

71. Beckerman H, Roebroeck ME, Lankhorst GJ, Becher JG, Bezemer PD, Verbeek AL. Smallest real difference, a link between reproducibility and responsiveness. Qual Life Res Int J Qual Life Asp Treat Care Rehabil. 2001;10(7):571–8.

72. Jacobson NS, Truax P. Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. J Consult Clin Psychol. 1991 Feb;59(1):12–9.

73. Fitzpatrick R, Norquist JM, Jenkinson C. Distribution-based criteria for change in health-related quality of life in Parkinson's disease. J Clin Epidemiol. 2004;57(1):40–4.

74. Guyatt GH, Bombardier C, Tugwell PX. Measuring disease-specific quality of life in clinical trials. Can Med Assoc J. 1986;134(8):889–95.

75. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. J Chronic Dis. 1987;40(2):171–8.

76. Wyrwich KW, Norquist JM, Lenderking WR, Acaster S, Industry Advisory Committee of International Society for Quality of Life Research (ISOQOL). Methods for interpreting change over time in patient-reported outcome measures. Qual Life Res. 2013;22(3):475–83.

77. Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 1989;10(4):407–15.

78. Wyrwich KW, Metz SM, Kroenke K, Tierney WM, Babu AN, Wolinsky FD. Triangulating patient and clinician perspectives on clinically important differences in health-related quality of life among patients with heart disease. Health Serv Res. 2007;42(6):2257–74.

79. Guyatt GH, Osoba D, Wu AW, Wyrwich KW, Norman GR. Methods to explain the clinical significance of health status measures. Mayo Clin Proc. 2002;77(4):371–83.

80. Barrett B, Brown D, Mundt M, Brown R. Sufficiently important difference: expanding the framework of clinical significance. Med Decis Making. 2005;25(3):250–61.

81. Schünemann HJ, Akl EA, Guyatt GH. Interpreting the results of patient reported outcome measures in clinical trials: the clinician's perspective. Health Qual Life Outcomes. 2006;4:62.

82. Ringash J, O'Sullivan B, Bezjak A, Redelmeier DA. Interpreting clinically significant changes in patient-reported outcomes. Cancer. 2007 Jul 1;110(1):196–202.

83. Wyrwich KW, Wolinsky FD. Identifying meaningful intra-individual change standards for health-related quality of life measures. J Eval Clin Pract. 2000;6(1):39–49.

84. Sloan JA, Cella D, Hays RD. Clinical significance of patient-reported questionnaire data: another step toward consensus. J Clin Epidemiol. 2005;58(12):1217–9.

85. Terwee CB, Roorda LD, Knol DL, De Boer MR, De Vet HCW. Linking measurement error to minimal important change of patient-reported outcomes. J Clin Epidemiol. 2009;62(10):1062–7.

86. Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. J Chronic Dis. 1986;39(11):897–906.

87. Behl P, Stefurak TL, Black SE. Progress in clinical neurosciences: cognitive markers of progression in Alzheimer's disease. Can J Neurol Sci J Can Sci Neurol. 2005 May;32(2):140–51.

88. Martinez-Martin P, Prieto L, Forjaz MJ. Longitudinal metric properties of disability rating scales for Parkinson's disease. Value Health. 2006;9(6):386–93.

89. Martínez-Martín P, Rodríguez-Blázquez C, Forjaz MJ, de Pedro J. The Clinical Impression of Severity Index for Parkinson's Disease: international validation study. Mov Disord. 2009;24(2):211–7.

90. Rodriguez-Blazquez C, Frades-Payo B, Forjaz MJ, de Pedro-Cuesta J, Martinez-Martin P. Psychometric attributes of the Hospital Anxiety and Depression Scale in Parkinson's disease. Mov Disord. 2009;24(4):519–25.

91. Rodriguez-Blazquez C, Forjaz MJ, Frades-Payo B, de Pedro-Cuesta J, Martinez-Martin P. Independent validation of the scales for outcomes in Parkinson's disease-autonomic (SCOPA-AUT). Eur J Neurol. 2010;17(2):194–201.

92. Virués-Ortega J, Rodríguez-Blázquez C, Micheli F, Carod-Artal FJ, Serrano-Dueñas M, Martínez-Martín P. Cross-cultural evaluation of the modified Parkinson Psychosis Rating Scale across disease stages. Mov Disord. 2010;25(10):1391–8.

93. ELEP Group. [A longitudinal study of patients with Parkinson's disease (ELEP): aims and methodology]. Rev Neurol. 2006;42(6):360–5.

94. Martinez-Martin P, Rodriguez-Blazquez C, Abe K, Bhattacharyya KB, Bloem BR, Carod-Artal FJ, et al. International study on the psychometric attributes of the non-motor symptoms scale in Parkinson disease. Neurology. 2009;73(19):1584–91.

95. Martinez-Martin P, Rodriguez-Blazquez C, Alvarez-Sanchez M, Arakaki T, Bergareche-Yarza A, Chade A, et al. Expanded and independent validation of the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS). J Neurol. 2013;260(1):228–36.

96. Martínez-Martín P, Forjaz MJ, Cubo E, Frades B, de Pedro Cuesta J. Global versus factor-related impression of severity in Parkinson's disease: A new clinimetric index (CISI-PD). Mov Disord. 2006;21(2):208–14.

97. Friedberg G, Zoldan J, Weizman A, Melamed E. Parkinson Psychosis Rating Scale: a practical instrument for grading psychosis in Parkinson's disease. Clin Neuropharmacol. 1998;21(5):280–4.

98. The Unified Parkinson's Disease Rating Scale (UPDRS): status and recommendations. Mov Disord. 2003;18(7):738–50.

99. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. Acta Psychiatr Scand. 1983;67(6):361–70.

100. Wild D, Grove A, Martin M, Eremenco S, McElroy S, Verjee-Lorenz A, et al. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. Value Health. 2005;8(2):94–104.

101. Horn JL. A rationale and test for the number of factors in factor analysis. Psychometrika. 1965;30:179–85.

102. Beaton DE, Bombardier C, Katz JN, Wright JG. A taxonomy for responsiveness. J Clin Epidemiol. 2001;54(12):1204–17.

103. Movement Disorders Society. MDS Rating Scales [Internet]. [cited 2014 Jul 10]. Available from: http://www.movementdisorders.org/MDS/Education/Rating-Scales.htm

104. Schrag A, Barone P, Brown RG, Leentjens AFG, McDonald WM, Starkstein S, et al. Depression rating scales in Parkinson's disease: Critique and recommendations. Mov Disord. 2007;22(8):1077–92.

105. Leentjens AFG, Dujardin K, Marsh L, Richard IH, Starkstein SE, Martinez-Martin P. Anxiety rating scales in Parkinson's disease: a validation study of the Hamilton anxiety rating scale, the Beck anxiety inventory, and the hospital anxiety and depression scale. Mov Disord. 2011;26(3):407–15.

106. Pavy-Le Traon A, Amarenco G, Duerr S, Kaufmann H, Lahrmann H, Shaftman SR, et al. The Movement Disorders task force review of dysautonomia rating scales in Parkinson's disease with regard to symptoms of orthostatic hypotension. Mov Disord. 2011;26(11):1985–92.

107. Evatt ML, Chaudhuri KR, Chou KL, Cubo E, Hinson V, Kompoliti K, et al. Dysautonomia rating scales in Parkinson's disease: sialorrhea, dysphagia, and constipation--critique and recommendations by movement disorders task force on rating scales for Parkinson's disease. Mov Disord. 2009;24(5):635–46.

108. Fernandez HH, Aarsland D, Fénelon G, Friedman JH, Marsh L, Tröster AI, et al. Scales to assess psychosis in Parkinson's disease: Critique and recommendations. Mov Disord. 2008;23(4):484–500.

109. Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, et al. Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord. 2008;23(15):2129–70.

110. Goetz CG. Unified Parkinson's Disease Rating Scale (UPDRS) and Movement Disorders Society Revision of the UPDRS (MDS-UPDRS). In: Sampaio C, Goetz CG, Schrag A, editors. Rating scales in Parkinson's disease. New York: Oxford University Press; 2012. p. 62–83.

111. Jankovic J. Parkinson's disease: clinical features and diagnosis. J Neurol Neurosurg Psychiatry. 2008;79(4):368–76.

112. Maetzler W, Liepelt I, Berg D. Progression of Parkinson's disease in the clinical phase: potential markers. Lancet Neurol. 2009;8(12):1158–71.

113. Eggers C, Pedrosa DJ, Kahraman D, Maier F, Lewis CJ, Fink GR, et al. Parkinson subtypes progress differently in clinical course and imaging pattern. PloS One. 2012;7(10):e46813.

114. Marras C, Lang A. Parkinson's disease subtypes: lost in translation? J Neurol Neurosurg Psychiatry. 2013;84(4):409–15.

115. Szewczyk-Krolikowski K, Tomlinson P, Nithi K, Wade-Martins R, Talbot K, Ben-Shlomo Y, et al. The influence of age and gender on motor and non-motor features of early Parkinson's disease: initial findings from the Oxford Parkinson Disease Center (OPDC) discovery cohort. Parkinsonism Relat Disord. 2014;20(1):99–105.

116. Leentjens AFG, Dujardin K, Pontone GM, Starkstein SE, Weintraub D, Martinez-Martin P. The Parkinson Anxiety Scale (PAS): development and validation of a new anxiety scale. Mov Disord. 2014;29(8):1035–43.

117. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care. 2003;41(5):582–92.

118. De Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. Health Qual Life Outcomes. 2006;4:54.

119. Chaudhuri KR, Rojo JM, Schapira AHV, Brooks DJ, Stocchi F, Odin P, et al. A proposal for a comprehensive grading of Parkinson's disease severity combining motor and non-motor assessments: meeting an unmet need. PloS One. 2013;8(2):e57221.

120. Lee CS, Schulzer M, Mak EK, Snow BJ, Tsui JK, Calne S, Hammerstad J, Calne DB. Clinical observations on the rate of progression of idiopathic parkinsonism. Brain. 1994;117(Pt 3):501-7.

121. Poewe W, Mahlknecht P. The clinical progression of Parkinson's disease. Parkinsonism Relat Disord. 2009;15 Suppl 4:S28-32.

122. Poewe W. Clinical measures of progression in Parkinson's disease. Mov Disord. 2009;24 Suppl 2:S671-6.

123. Bugalho P, Viana-Baptista M. Predictors of cognitive decline in the early stages of Parkinson's disease: a brief cognitive assessment longitudinal study. Park Dis. 2013;2013:912037.

124. Vu TC, Nutt JG, Holford NHG. Progression of motor and nonmotor features of Parkinson's disease and their response to treatment. Br J Clin Pharmacol. 2012;74(2):267–83.

125. Honig H, Antonini A, Martinez-Martin P, Forgacs I, Faye GC, Fox T, et al. Intrajejunal levodopa infusion in Parkinson's disease: a pilot multicenter study of effects on nonmotor symptoms and quality of life. Mov Disord. 2009;24(10):1468–74.

126. Martinez-Martin P, Reddy P, Antonini A, Henriksen T, Katzenschlager R, Odin P, et al. Chronic Subcutaneous Infusion Therapy with Apomorphine in Advanced Parkinson's Disease Compared to Conventional Therapy: A Real Life Study of Non Motor Effect. J Park Dis. 2011;1(2):197–203.

127. Reddy P, Martinez-Martin P, Rizos A, Martin A, Faye GC, Forgacs I, et al. Intrajejunal levodopa versus conventional therapy in Parkinson disease: motor and nonmotor effects. Clin Neuropharmacol. 2012;35(5):205–7.

128. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol. 2000;53(5):459–68.

129. Norman GR, Wyrwich KW, Patrick DL. The mathematical relationship among different forms of responsiveness coefficients. Qual Life Res. 2007;16(5):815–22.

130. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. Med Care. 1989 Mar;27(3 Suppl):S178–89.

131. Turner D, Schünemann HJ, Griffith LE, Beaton DE, Griffiths AM, Critch JN, et al. The minimal detectable change cannot reliably replace the minimal important difference. J Clin Epidemiol. 2010;63(1):28–36.

132. Martinez-Martin P, Kurtis MM. Health-related quality of life as an outcome variable in Parkinson's disease. Ther Adv Neurol Disord. 2012;5(2):105–17.